Verbal Reasoning

Thad A. Polk

August 31, 1992
CMU-CS-92-178

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Submitted in partial fulfillment of the requirements for an Interdisciplinary doctoral degree in Computer Science and Psychology from Carnegie Mellon University

This research was supported in part by a fellowship from the Eastman Kodak Company, by the James S. McDonnell Foundation under Grant No. 91-34, by the Office of Naval Research under contract number N00013-91-J-1527, and by the Avionics Laboratory, Wright Research and Development Center, Aeronautical Systems Division (AFSC), U.S. Air Force, Wright-Patterson AFB, Ohio 45433-6543 under Contract F33615-90-C-1465, ARPA Order No. 7597.

The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of Kodak, the James S. McDonnell Foundation, ONR, DARPA or the U.S. government.

Keywords: Verbal reasoning, deduction, inference, syllogisms.

Carnegie Mellon

School of Computer Science

DOCTORAL THESIS
in the field of
Computer Science and Psychology

Verbal Reasoning

THAD A. POLK

Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

ACCEPTED:

Kurt A. Van L
THESIS COMMITTEE CHAIR — Aug 31, 1992 DATE

DEPARTMENT HEAD — 9/4/92 DATE

APPROVED:

SCS DEAN — 9/2/92 DATE

H&SS DEAN — 9-?-92 DATE

Table of Contents

List of Figures

Abstract

Previous theories of human reasoning have all been based on what might be called the transduction paradigm: people *encode* the problem statement into an internal representation, *reason* using processes devoted specifically to that purpose, and then *decode* the result. The nature of the intermediate reasoning process has been a major source of debate among cognitive psychologists. Some researchers have proposed that this process can best be characterized as the application of formal rules of inference while others have argued that it corresponds to a search for alternative mental models of the problem statement. I believe that the transduction paradigm itself is in error, at least for the standard tasks that have been used in studying human deductive reasoning. My thesis is that most untrained subjects lack sophisticated reasoning-specific mechanisms for solving these tasks, and that, in their absence, they attempt to make progress by repeatedly applying linguistic processes to encode and reencode the problem statement. I refer to this type of behavior as verbal reasoning.

The idea of verbal reasoning arose out of VR, a computational model of human behavior on categorical syllogisms. It simulates human behavior on every variant of the syllogism task using purely linguistic processes. VR models *all* of the standard phenomena that have been discovered about syllogistic reasoning and makes a number of novel predictions that have been empirically confirmed. Furthermore, using a set of individual difference parameters, it has been tailored to fit the behavior of individual subjects with an accuracy that rivals the test-retest reliability of the subjects themselves. Since VR only uses linguistic processes, it provides compelling evidence that human syllogistic reasoning can best be characterized as verbal reasoning. Finally, to show that verbal reasoning generalizes to other tasks, I analyzed Johnson-Laird and Byrne's recent attempt to provide accounts of behavior on all the standard deductive reasoning tasks. In most cases, their explanations only depend on assumptions about how the problem statement is comprehended and thus constitute verbal reasoning accounts. In the few cases in which their explanations depend on reasoning-specific mechanisms, verbal reasoning provides a more parsimonious account.

Thesis committee:

Allen Newell, former chairman
Kurt VanLehn, acting chairman
Robert Siegler
Philip Johnson-Laird, Princeton University

Preface

Like most dissertations, this one is based on collaborative work between a student and his advisor. In particular, chapters 2 through 7 are based on an article (Polk & Newell, 1992) that Allen Newell and I authored together. Rather than changing all the first-person plural pronouns in that paper to be first-person singular (which I felt would be inappropriate), I decided to leave them unchanged in the dissertation. I hope that these references will not cause too much confusion.

Acknowledgements

Outside of my immediate family, the people that I met during graduate school have probably influenced me more than anyone I've ever known. I can honestly say that because of my relationships with them I have grown more professionally and personally during these six years than at any other time in my life. I'm really grateful for this opportunity to thank some of them.

The first person I want to thank is my advisor, Allen Newell. His illness and recent death have obviously cast a shadow over the last few months of my graduate work. But I'm extremely thankful that I had the opportunity to work with him so closely over the last few years. A couple of years ago Allen confessed to me that I was one the "greenest" graduate students he had ever had when I first started. I was a mathematics major as an undergraduate and had absolutely no knowledge of cognitive psychology or artificial intelligence. The fields just seemed glamorous so I wanted to give them a try. When I think back to my first scientific talk in the fall of 1986, I can't help but cringe in embarrassment. As I hope is evidenced by this thesis, I've come a long way since then and that's really a tribute to Allen, not to me. Despite the fact that he was a world-reknown researcher and I was just an ignorant graduate student playing at being a scientist, he took a genuine interest in my development and treated me with the utmost respect. He taught me virtually everything I know about science and being a scientist. I'll miss him more than I can say.

I also want to thank my wife Norma. Quite simply, she is the most important person in the world to me. Our first year of marriage has been the happiest and most fulfilling time of my life. I only wish we had gotten married sooner! Her constant support and affection have made the last year of my graduate work not only bearable, but genuinely wonderful. And now we're expecting our first baby at the end of February! It puts everything else into perspective.

Of course, my parents have had more of an impact on my life than anyone else. They raised me, put me through school, and have supported me unconditionally throughout my life. I obviously wouldn't be finishing this dissertation if it weren't for everything they did.

I've made too many friends in Pittsburgh to name them all. Let me just collectively thank the Soar group, my present and past officemates, and everyone else in this wonderful community who have been such great friends. I especially want to thank Rick Lewis who has been everything a friend could be. We've lived together, worked together, and played together over the last five years and moving away from Rick is one of my greatest regrets in leaving Pittsburgh. Who am I going to play golf, croquet, etc. with? I also want to mention Anurag, Tony, Alonso, and Milind with whom I've had many great times. Thanks to Kurt Vanlehn, Phil Johnson-Laird, and Bob Siegler (my thesis committee) and to the members of my degree committee (Martha Farah, Bonnie John, Pat Carpenter, and Jaime Carbonell) for their contributions to my research and course work. And a special thanks to everyone at the Church of the Ascension, especially the house group and choir, for their friendship and support. Through their prayers, God helped me through this degree.

Chapter 1
Introduction

Human beings are constantly faced with the problem of inferring something novel from available information. In daily life, we often solve this problem effortlessly and without a second thought. If my friend tells me he will either be at home or at work and I cannot reach him at work, I fully expect him to pick up the phone when I call his house. I come to this conclusion so quickly and easily it hardly seems like reasoning. Yet I could only reach it by combining the information from at least two separate assumptions — (1) my friend is either at home or at work and (2) my friend is not at work (not to mention the assumption that my friend told the truth, that the phones are working, ...). Neither assumption alone would lead me to the conviction that my friend is at home, so I must have put them together in some way. How did I do it? More specifically, what cognitive processes did I apply so effortlessly to reach that conclusion?

The goal of this dissertation is to provide some insight into this kind of question. I begin with a specific reasoning task, the categorical syllogism, and in Chapter 2 describe the task, some of the empirical results surrounding it, and the approaches that have been taken to studying it. In Chapter 3 I present VR, a computational model of human behavior on categorical syllogisms. In the next chapter I show how VR can be adapted to all the standard variants of the syllogism task. In chapters 5 and 6, I go on to show that VR can successfully model both aggregate data and the behavior of individual subjects[1]. Chapter 7 steps back and discusses the general view of reasoning suggested by VR, namely verbal reasoning, and how it compares with other theories. Finally, in Chapter 8 I show that verbal reasoning generalizes beyond categorical syllogisms to all the standard deductive reasoning tasks.

[1] Thanks to Phil Johnson-Laird for graciously providing us with the raw data from his experiments with Bruno Bara and Mark Steedman which we used in these analyses.

Chapter 2
Categorical Syllogisms

The study of reasoning has been going on almost since the beginning of experimental psychology. The typical approach has been to choose a relatively straightforward task that requires reasoning and then to study human behavior on that task. Errors are particularly interesting since they help to suggest and constrain the underlying processes that subjects are using. Recurring error patterns and other behavioral regularities are most interesting of all since they suggest characteristics of the underlying processes that could generalize across many subjects. Consequently, scientists have sought out and studied reasoning tasks that are relatively hard (so that subjects make a significant number of errors) and that lead to a number of robust regularities in behavior. A good example of such a task is the categorical syllogism.

2.1. The categorical syllogism task

Categorical syllogisms are reasoning problems consisting of two premises and a conclusion (Figure 2-1, left). Each premise relates two terms (x and y) in one of four ways (Figure 2-1, middle), and they share a common *middle term* (*bowlers* on the left of the figure). A conclusion states a relation between the two terms that are not common (the *end terms — archers* and *chefs*), but valid conclusions need not exist (in which case the correct response is that there is no valid conclusion — abbreviated NVC). The three terms x,y,z can occur in four different orders (called *figures* — Figure 2-1, right), producing 64 distinct premise pairs. Different versions of the task require determining the validity of each member of a set of conclusions (Wilkins, 1928; Sells, 1936), choosing a valid conclusion from a set of alternatives (Chapman & Chapman, 1959; Ceraso & Provitera, 1971; Revlis, 1975b; Dickstein, 1975), evaluating the validity of a given conclusion (Janis & Frick, 1943), and generating a valid conclusion given only the premises (Johnson-Laird & Bara, 1984).

Premise 1: No archers are bowlers.
Premise 2: Some bowlers are chefs.
Conclusion: Some chefs are not archers.

A: All x are y.
I: Some x are y.
E: No x are y.
O: Some x are not y.

P1: Axy P2: Iyz	P1: Oyx P2: Ayz
P1: Exy P2: Ezy	P1: Iyx P2: Ozy

Figure 2-1: Categorical syllogism task.

2.2. Regularities in syllogistic reasoning

Some categorical syllogisms are extremely easy (e.g., *All x are y. All y are z. Therefore, what necessarily follows?*) while others are beyond the competence of most untrained subjects (e.g., *Some x are y. No z are y. Therefore, what necessarily follows?*)[2]. Overall, many undergraduates correctly solve less than half of the 64 problems that require generating a conclusion. so there are plenty of errors to analyze. Furthermore, when responses are aggregated over groups of subjects, some striking regularities emerge. Figure 2-2 lists some of the most important. We chose these seven because (1) they are robust (i.e., they have been replicated), (2) they have been cited by more than one author, and (3) they can be described without reference to any theory. Wilkins (1928) and Sells (1936) have also claimed evidence for a *concreteness effect* — that subjects are more accurate on syllogisms involving concrete (archer, bowler) rather than abstract (A, B, C) terms. The observed effect was not particularly large (around 10% fewer errors in both data sets) and, to our knowledge, has not been replicated. Indeed, Gonzales-Marques (1985) tried to do so and failed[3]. Furthermore, Revlis and colleagues (1975a,1978) have argued that when believability effects are controlled (specifically, when the converse of each premise does not contradict beliefs), the effect disappears. If so, then the Wilkins and Sells results may have been the result of belief bias rather than a concreteness effect. In any case, the effect is not well established and so we have not included it in Figure 2-2. Other empirical results have also been found, but they have typically been cited in favor of a single theory, are often described in terms of that theory, and have usually not been replicated. Since most subjects exhibit the behavioral patterns of Figure 2-2, a theory that explains them should generalize across many subjects. In general, the subjects have been college students.

[2] In one experiment (Johnson-Laird & Bara, 1984), 19/20 subjects correctly solved the first syllogism (*All x are z*), while 0/20 solved the second (Some x are not z).

[3] This paper is in Spanish. The assertion is based on an English translation of the abstract.

1. **Difficulty** — The average subject makes many errors (often making errors on approximately half the tasks) (Wilkins, 1928; Dickstein, 1975).
2. **Validity effect** — The average subject does better than chance (Revlis, 1975b; Johnson-Laird & Steedman, 1978).
3. **Atmosphere effect** — Excluding NVC responses, (1) if either premise is negative (*No x are y* or *Some x are not y*), most responses are negative, otherwise most are positive; and (2) if either premise is particular (referring to a subset — *Some x are y* or *Some x are not y*), most responses are particular, otherwise most are universal (*All x are z* or *No x are z*) (Woodworth & Sells, 1935; Sells, 1936).
4. **Conversion effect** — Excluding NVC responses, many erroneous responses would be correct if the converse of one or both premises were assumed to be true (Chapman & Chapman, 1959; Revlis, 1975b).
5. **Figural effect** — Excluding NVC responses, if only one end term (x or z) appears as the subject of a premise, that term tends to appear as the subject of the conclusion (Johnson-Laird & Steedman, 1978; Johnson-Laird & Bara, 1984).
6. **Belief bias** — Subjects are more likely to generate and accept as valid a conclusion that they believe to be true rather one they believe to be false, independent of its true logical status (Wilkins, 1928; Janis & Frick, 1943; Morgan & Morton, 1944; Evans, Barston & Pollard, 1983; Oakhill & Johnson-Laird, 1985; Oakhill, Johnson-Laird & Garnham, 1989).
7. **Elaboration effect** — Subjects are more accurate if the premises are elaborated to be unambiguous (e.g., *All A are B, but Some B are not A*) (Ceraso & Provitera, 1971).

Figure 2-2: Regularities on categorical syllogisms.

2.3. Approaches to studying syllogistic reasoning

Categorical syllogisms would seem to be ideal tasks for studying human reasoning. They are relatively straightforward to understand and simple to administer, they are hard enough that untrained subjects make numerous errors, and they lead to numerous behavioral regularities. Perhaps as a result, there are over 150 journal articles related to syllogisms and numerous theories have been proposed. One of the first theories was the *atmosphere hypothesis* proposed by Woodworth & Sells (1935) which suggested that the type of relationships specified by the premises (their *mood*) created an atmosphere that led to erroneous conclusions. Compared with more recent theories, the atmosphere hypothesis is more descriptive than explanatory, and indeed, we have included its account as a regularity in Figure 2-2. Another early theory (Chapman & Chapman, 1959; also Revlis, 1975b) argued that many errors could be explained by assuming that subjects had illicitly converted one (or both) of the premises — assuming that *All x are y* implies *All y are x* and that *Some x are not y* implies *Some y are not x*. This hypothesis can also be stated as an empirical regularity in the data (thc conversion effect in Figure 2-2). More recently, several authors (Erickson, 1974; Guyote & Sternberg, 1981; Fisher, 1981) have proposed that people solve syllogisms using representations analogous to Euler circles. In this notation, each class of individuals is represented by a separate circle. Overlapping parts of circles represent individuals that are in more than one class. Venn diagrams are similar except all circles are drawn as overlapping initially and some notation (e.g., shading) is used to indicate classes of individuals that are known to be present or absent. Finally, Johnson-Laird (Johnson-Laird, 1983; Johnson-Laird & Bara, 1984; Johnson-Laird & Byrne, 1991) has presented a theory based on the idea that people construct and manipulate *mental models* of what the premises describe.

Our approach to studying syllogistic reasoning has been to construct and then repeatedly evaluate and refine computational models for the task. These models serve as operational theories and can be run to generate predictions. Until recently our models were all built in the Soar architecture which implements problem spaces in terms of a production system (a recognition memory) (Polk & Newell, 1988; Polk, Newell & Lewis, 1989). A major strength of production systems is that they make it possible to model the dynamic flow of control characteristic of human behavior. In most programming languages, the sequence of actions is specified explicitly in the program itself (first perform the action specified by the first line, then perform the action specified by the second line, ..., with some conditionals and jumps). In production systems, a set of IF-THEN rules specifies what to do in various situations. When executed, these rules *dynamically control the* sequence of actions taken by the program. While such a dynamic control structure is desirable when modeling human behavior, it comes with a price — production systems tend to run significantly slower than other programming languages. Since some of the analyses we will present later in this dissertation require running the model a very large number of times, we decided to build our most recent model — VR(Syl), hereafter just VR (for verbal reasoner) — in Lisp. Consequently, we had to specify the flow of control ahead of time (essentially building in a flowchart) and this obscures the dynamic flow of control that is so characteristic of human behavior. So while VR has a built-in control structure, this is a characteristic of the programming environment and we do not want to attribute it to our subjects.

Chapter 3

VR: A Computational Model for Categorical Syllogisms

Figure 3-1 presents the general control structure of VR when generating a conclusion from a pair of premises. In the next chapter we will describe how VR can be adapted to other variants of the task.

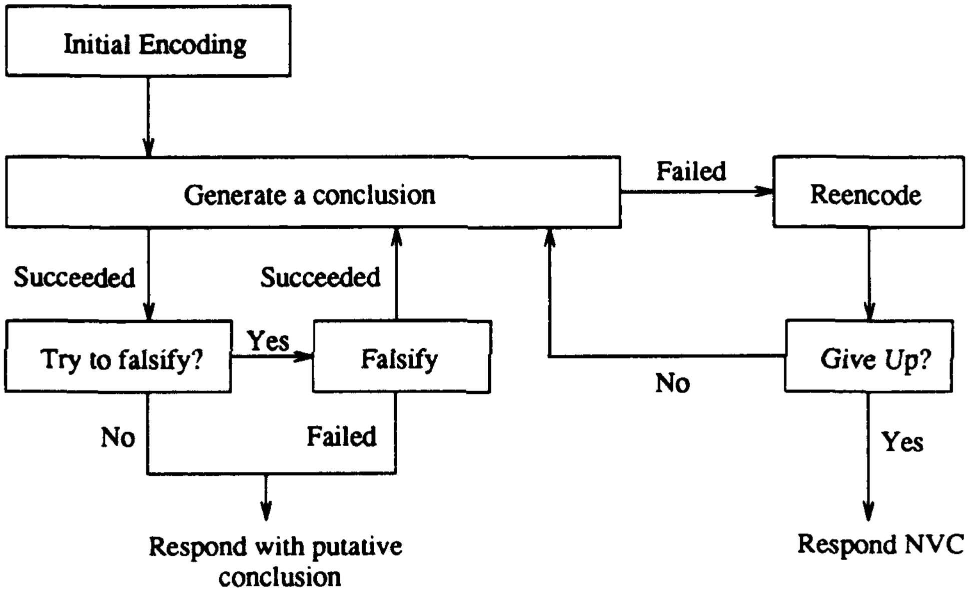


Figure 3-1: Control structure of VR.

At the most general level, VR's structure reflects the basic demands of the task. At the very least, the task requires encoding the premises (initial encoding) and producing a response (generate a conclusion). If VR fails to successfully generate a conclusion, then it repeatedly reencodes the premises until generation succeeds or until it gives up. Reencoding in this context is a natural response to the task demand of producing a conclusion. After all, the system lacks the knowledge to relate the end terms, the main source of knowledge for the problem is the premises, and encoding is the most natural way of extracting knowledge from a premise. If repeated reencoding fails to lead to a conclusion, then VR gives up and assumes that there is no valid conclusion (NVC). If it succeeds in generating a conclusion, then it may or may not try to falsify it (controlled by a parameter in the model). Falsification is an attempt by the system to respond to the task demand of producing a *valid* conclusion. If the system knows that valid conclusions cannot be falsified, then trying to do so is obviously appropriate. On the other hand, if the system lacks any knowledge about how to verify the validity of a conclusion, then a rational alternative is to produce its putative conclusion as the response — it certainly has no reason to prefer any other response. We now turn to specifying the details of each of the major processes in Figure 3-1: initial encoding, conclusion generation, reencoding, giving up, and falsification.

3.1. Initial encoding

When trying to solve a syllogism, the first thing VR does is to encode the premises. Following Johnson-Laird (Johnson-Laird, 1983; Johnson-Laird & Bara, 1984; Johnson-Laird & Byrne, 1991), VR tries to construct a *mental model* of a situation in which the premises are true. Mental models have been shown to be useful in explaining a wide variety of behavior in reasoning and language research. They consist of a set of objects with properties and relations among them. The defining characteristic of mental models is that they satisfy the *structure-correspondence principle* — all objects, properties, and relations in the mental model must map 1-1 into objects, properties, and relations in the situation being represented (the referent). In addition to their empirical success and intuitive appeal, there are computational reasons to believe humans must be using mental models — they can be processed in bounded time. As Newell (1990) noted, structure-correspondence implies that inexpensive match-like and counting processes can be used to process mental models. For example, any two mental models that refer to the same situation must themselves correspond. Consequently, determining whether two mental models have the same referent is simply a matter of putting the two in 1-1 correspondence and checking for inconsistencies. Such a process is bounded by the size of the representation. Representations that do not satisfy the structure-correspondence principle have no such bound. For example, determining whether two sets of formulas in first-order logic refer to the same situation can take an arbitrary amount of time since there is no inherent bound on the number of proof steps required[4]. The bounded cost of processing mental models comes with a price, however — limited expressive power. Since all objects must map into objects in the referent, pure mental models cannot contain quantifiers, implication symbols, or explicit disjunction[5]. Even negation violates the

[4] To take just one example, consider the following two sets of formulas: $\{A(1), A(n) \rightarrow A(n+1)\}$ and $\{\neg A(k)\}$. In the first set, all natural numbers satisfy predicate A, but in the second the constant k does not satisfy A. Proving they represent different situations requires $k-1$ induction steps (assuming the calculus does not provide one-step induction). Since k is arbitrary, so is the amount of computation.

[5] Since models typically only represent a subset of the objects, properties and relations in the referent, a given model may be consistent with an infinite number of referents (that differ in ways not made explicit in the model). In this sense, a model can represent disjunction. The point is that it is impossible, within a single pure model, to represent *explicitly* a set of alternative situations.

structure-correspondence principle. In order to provide more expressive power without sacrificing bounded processing, VR uses an *annotated model* representation. Specifically, VR allows properties to be annotated with a **not** flag indicating that the specified property does not exist in the referent. Such an annotation violates structure-correspondence, but since its effect is limited to a known property and object, the processing cost remains bounded. Other useful annotations may exist, but are not part of VR.

After encoding a proposition, the resulting annotated model is guaranteed to represent a situation in which that proposition is true. But, in general, there are an infinite variety of such situations, and most of them are unwarranted or incomplete with respect to the initial proposition. That is, an annotated model may encode information that is not inherent in a proposition (and be unwarranted) or fail to encode information that is (and be incomplete). Figure 3-2 gives examples of each. On the left is an annotated model that is unwarranted with respect to the proposition *Some archers are bowlers.* The annotated model consists of two objects represented as a set of properties enclosed in parentheses. The first object is an archer and bowler, while the second is an archer and explicitly *not* a bowler. The proposition *Some archers are bowlers* is true in this annotated model, but in addition the model encodes information that is unwarranted given only that proposition — the fact that some archers are *not* bowlers. On the right is an example of an incomplete annotated model. Once again, the model supports the initial proposition (*No archers are bowlers* in this case), but now it fails to reflect the fact that the second object (the bowler) cannot be an archer. In this case then, the annotated model is incomplete with respect to the initial proposition.

Unwarranted Annotated Model

Proposition: **Some archers are bowlers.**

Annotated Model: **(archer bowler)**
(archer -bowler)

Incomplete Annotated Model

Proposition: **No archers are bowlers.**

Annotated Model: **(archer -bowler)**
(bowler)

Figure 3-2: Examples of unwarranted and incomplete annotated models.

In Figure 3-2, the properties for each model object have equal status. They form an unordered set and none is easier to access than the others. In the top object in the model on the left, for example, there is no distinction between the properties archer and bowler. But since the proposition itself was *about* archers and not bowlers, it seems questionable that the two properties should have equal status. The assumption appears even more unlikely for the bottom object in which archer has the same status as the negated property -bowler. The annotated model representation in VR distinguishes *identifying properties* (such as archer above) from *secondary properties* (like bowler). For syllogisms, the identifying properties simply correspond to the topics of the propositions being encoded. These properties are accessed before others during VR's processing. Specifically, when VR tries to generate conclusions based on its annotated models, it tries conclusions about identifying properties before trying conclusions about secondary properties. In the rest of this dissertation, identifying properties will be indicated by the ' symbol. For example, (archer' bowler) is an annotated model object with identifying property archer and secondary property bowler.

The model on the right of Figure 3-2 is incomplete because it does not reflect the fact that no bowler could be an archer. But where is this knowledge about bowlers? Obviously, it is in the proposition *No archers are bowlers*, even though that statement is about archers, not bowlers. So this proposition does contain some information about its non-topic property (bowler), but this information is less accessible than is the information about its topic (archers). In keeping with this distinction, VR only extracts *direct knowledge* (about the topic) during its initial encoding. For example, suppose encoding the first premise of a syllogism leads to an annotated model with a single object (A' B). Then an initial encoding of *No C are B* will not affect that object since the direct knowledge is about C and that model object does not have property C. It is only if this direct knowledge is insufficient to produce a conclusion that reencoding will try to extract indirect knowledge (about non-topic properties).

When VR tries to falsify a conclusion, it needs to be able to detect inconsistencies between a proposition and an annotated model. More generally, it needs to extract *metaknowledge* about the relationship between the proposition and the annotated model. Humans exhibit this ability all the time in normal conversation when they notice that a proposition is redundant or inconsistent with previous statements. Rather than having a separate process to extract metaknowledge, VR's encoding process itself provides this capability. Specifically, when encoding a proposition into an annotated model, VR also flags whether the proposition was *redundant*, *inconsistent*, or *irrelevant* with respect to the annotated model. The choice of flag is based on how the annotated model was affected during encoding: if it was unchanged then the proposition was redundant, if an existing property was removed or replaced then the proposition was inconsistent, and if the existing objects were unchanged but new unrelated objects were created that share no properties with the old objects then the proposition was irrelevant.

These assumptions about the representation (annotated models with identifying properties) and knowledge (only direct knowledge and some metaknowledge) produced by encoding provide important constraints, but there is still a wide range of representations consistent with these assumptions that could be produced. Indeed, we believe the way the premises are encoded is the major source of individual differences in syllogisms and later on we will try to address this diversity. For now, we will simply present a set of default encodings consistent with the above assumptions so that we can derive some predictions from VR. Figure 3-3 presents these default encodings.

For each of the four premise types the figure indicates how encoding that premise modifies the annotated model. For the universal premises (*All* and *No*) the default encoding is straightforward — all objects with property X are augmented with property Y (or -Y) and the X properties are marked as identifying. If there are no objects with property X, then a new object is created. The default encodings for the particular premises (*Some* and *Some not*) consist of two parts corresponding to "Some x are (not) y" (the upper conditional sequences in the figure) and "other x may or may not be y" (the lower conditional sequences in the figure). As an example, consider reading *Some B are C* when the annotated model contains a single object with properties A and B — (A' B). Following the upper conditional sequence in the figure, there is no object with properties B & C, but the object (A' B) does have property B without property -C. So VR augments the most recently accessed (MR in Figure 3-3) such object (there is only one in this case) with property C and marks B as identifying. The resulting object has all three properties — (A' B' C). Similarly, following the lower conditional sequence, the initial object (A' B) has property B, but has neither C nor -C. Consequently, VR just marks B as identifying in the most recently accessed (MR) such object (again there is only one). Notice that the same initial object is used in both cases leading to two different objects after the encoding. Combining these two subencodings leads to the following annotated model — (A' B' C) and (A' B'). This interpretation of the quantifier "some" basically corresponds to "some but not necessarily all of the objects previously discussed". Other interpretations could easily be constructed ("some but not all of a different set of objects", etc.) by making appropriate modifications to the encodings in Figure 3-3. It should be noted that adding a property requires deleting its negation if present (this is the additional processing cost incurred by using negation).

3.2. Conclusion generation

Once VR has encoded the premises, it attempts to produce a conclusion based on its annotated model. It does so by generating simple propositions[6] and testing whether they are legal syllogism conclusions. As previously mentioned, VR tries propositions about identifying properties first. If none succeed, a parameter controls whether VR will try propositions about secondary properties. Associated with each simple proposition is a template that specifies the conditions under which that proposition is true in the annotated model. When the annotated model matches the template, then the associated simple proposition is proposed. All proposed propositions are then tested to see if they are legal syllogism conclusions. If there are none (including any about secondary properties if they were proposed), generation fails and reencoding is evoked. If there is at least one,

[6] These propositions satisfy a number of constraints: they never involve more than two terms, only the first term may be quantified, and they do not allow disjunction or conjunction.

All X are Y If there is an object with X Then X → X' in all such objects and augment them all with Y Else create new object (X' Y)	**No X are Y** If there is an object with X Then X → X' in all such objects and augment them all with -Y Else create new object (X' -Y)
Some X are Y If there is an object with X & Y Then X → X' in MR such object Else if there is an object with X & not -Y Then X → X' in MR such object and augment it with Y Else create new object (X' Y) If there is an object with X, not Y, & not -Y Then X → X' in MR such object Else create new object (X')	**Some X are not Y** If there is an object with X & -Y Then X → X' in MR such object Else if there is an object with X & not Y Then X → X' in MR such object and augment it with -Y Else create new object (X' -Y) If there is an object with X, not Y, & not -Y Then X → X' in MR such object Else create new object (X')

Figure 3-3: Default encodings for VR (MR = most recently accessed).

generation succeeds. In practice, there is rarely more than one proposed conclusion that is legal. When there is, VR produces the set of conclusions it considers equally probable[7].

Generation is much more tightly constrained than encoding, but the above assumptions still allow for some variety in the generation process — specifically, in the templates that trigger each proposition. Figure 3-4 presents a default set of such templates. Anytime the annotated model satisfies one of these templates, the associated conclusion is proposed.

3.3. Reencoding

If generation fails to produce a legal conclusion, VR tries to extract additional knowledge from the premises by reencoding. Unlike the initial encoding process which only extracts knowledge about a proposition's topic (direct knowledge), reencoding can extract indirect knowledge about non-topic properties. It first chooses a property based on the annotated model (the *reference property*), and then tries to extract additional knowledge about that property from any premise that mentions it (the *target proposition*). If the reference property is the topic of the target proposition, then reencoding works just like

[7] It could obviously choose randomly among the alternatives, but having the entire set allows us to compute the exact expected value of accurate predictions.

All X are Y Template: **There is an object with X'** **All objects with X have Y**	**Some X are Y** Template: **There is an object with X' and Y** **There is an object with X and not Y**
No X are Y Template: **There is an object with X'** **All objects with X have -Y**	**Some X are not Y** Template: **There is an object with X' and -Y** **There is an object with X and not -Y**

Figure 3-4: Default generation templates for VR.

initial encoding (extracting only direct knowledge). If, however, the reference property is not the target proposition's topic, then VR may extract indirect knowledge about that property from the proposition. For example, in trying to extract more knowledge about property B, VR might reencode the premise *No A are B* (assuming generation has failed so far). This premise does contain valid knowledge about B (that no B are A), and VR may or may not be able to extract it by reencoding that premise with respect to property B.

This description fails to specify whether VR will succeed in extracting indirect knowledge from a proposition and if so exactly what knowledge it will extract. We believe this is another major source of individual differences for this task and will return to it when we deal with individual data. For now, we will assume as a default that VR extracts no indirect knowledge at all. Given such an assumption, reencoding amounts to reapplying the initial encoding process since it can only extract direct knowledge. This fact does not imply that reencoding will be useless, however, since the results of initially encoding the second premise may influence the reencoding of the first. For example, consider the syllogism *All B are A, All C are B*. The second premise is about C, but encoding the first premise will not lead to any model objects with that property. Consequently, encoding the second premise will simply create a new object (C' B) and the model will not relate the end terms. But when the first premise is reencoded (even without extracting any indirect knowledge), the new model object will be augmented with property A so that the end terms are related.

There are at least three pieces of general evidence to recommend the view that reencoding can extract different knowledge from the same premise. First, as anyone who has studied categorical syllogisms knows, subjects reread the premises in the course of solving a problem. This behavior is difficult to account for assuming that all the information available in a proposition is extracted when it is first read. Second, subjects often reread the *same* premise many times strongly suggesting that they can extract different information from the same premise at different times. Finally, different subjects can exhibit very different reencoding behaviors. Accounting for such diversity is straightforward in VR since one can easily change what knowledge it will extract about a specific property from a given proposition. Indeed, this is one of the things we will do later in the dissertation when we use VR to explain individual differences.

The theory assumes that reference properties are chosen from among the properties in the annotated model. The goal of reencoding is to augment the annotated model such that it supports a valid conclusion, so extracting more information about a property that already appears in the model is a natural way to proceed. It certainly makes more sense than trying to get more information about a property that does *not* appear in the annotated model. But how should VR choose a reference property from among those that *are* in the model? The knowledge (if any) that people apply in making such a choice is far from clear. One rational heuristic is to prefer properties from objects that have been recently modified since these objects are less likely to reflect information from older propositions. This is the heuristic implemented in VR. Specifically, VR loops through the model objects starting with the most recently modified. For each object, it tries extracting more information about each property, from the newest to the oldest. As soon as the annotated model supports a conclusion (as soon as the model matches a generation template), generation is tried. If it succeeds, then VR goes on from there (either trying to falsify the conclusion or just producing it as its response). Otherwise, reencoding continues.

3.4. Giving up

VR has to stop reencoding at some point even if it has failed to generate a conclusion. When should it give up? The criteria that people use in making such a decision are not at all obvious. Nevertheless, it does seem clear that most subjects (at least the undergraduates that are typically studied) only give up as a last resort — when they feel confident that repeated attempts to extract more knowledge will not lead anywhere. VR reaches such a point after it has tried extracting more knowledge about every property in the annotated model, so this is when it decides to quit. It is at this point that VR responds with "No valid conclusion".

3.5. Falsification

As mentioned earlier, if VR succeeds in generating a legal conclusion, it may or may not try to falsify it (depending upon the value of a falsification parameter). Falsification is an attempt to find out whether there are situations in which the putative conclusion is false, but the premises are true. In VR, this corresponds to building an annotated model that contradicts the conclusion but supports the premises. The only means available to VR for constructing an annotated model is encoding and reencoding, so these are what it uses. Specifically, VR attempts to build an annotated model by encoding the premises and the *negation* of the putative conclusion. If constructing such an annotated model succeeds without running into any inconsistencies (recall that inconsistencies are extracted as metaknowledge during encoding), then the putative conclusion is known to be invalid and VR tries to produce a different conclusion. On the other hand, if VR encounters an inconsistency during this attempt, it assumes the putative conclusion to be valid and produces it as a response. In the default version of VR, the falsification parameter is set to "No" (i.e., do not try to falsify putative conclusions).

3.6. Summary of VR

Figure 3-5 summarizes VR by expanding Figure 3-1.

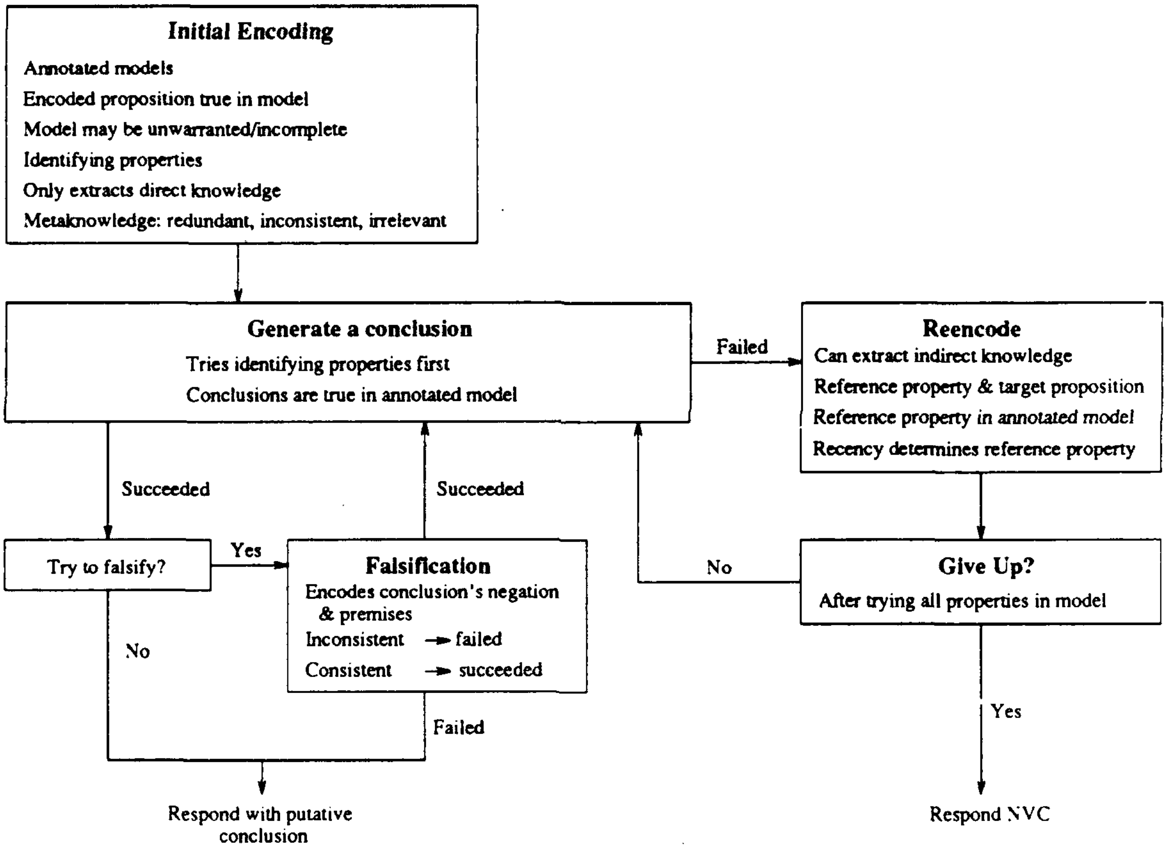

Initial Encoding
Annotated models
Encoded proposition true in model
Model may be unwarranted/incomplete
Identifying properties
Only extracts direct knowledge
Metaknowledge: redundant, inconsistent, irrelevant
Generate a conclusion
Tries identifying properties first
Conclusions are true in annotated model
Failed
Reencode
Can extract indirect knowledge
Reference property & target proposition
Reference property in annotated model
Recency determines reference property
Succeeded
Succeeded
Try to falsify?
Yes
Falsification
Encodes conclusion's negation & premises
Inconsistent → failed
Consistent → succeeded
No
Give Up?
After trying all properties in model
No
Failed
Yes
Respond with putative conclusion
Respond NVC

Figure 3-5: Summary of VR.

3.7. An example of VR's behavior

Figure 3-6 illustrates VR's behavior on the syllogism: *Some B are A. All B are C. Therefore, what necessarily follows?* In order to demonstrate falsification and the extraction of indirect knowledge, the default version of VR has been modified in two ways for this example: (1) the falsification parameter has been set to yes (so it attempts to falsify) and (2) this version of the system extracts *Some y are x* when reencoding *Some x are y* with respect to y.

VR's behavior on this syllogism follows the structure presented in Figures 3-1 and 3-5. After initially encoding the premises (step 1 in Figure 3-6), it repeatedly reencodes them with respect to different reference properties until it is able to generate a legal conclusion (steps 2-4). After falsification fails (steps 5-6), VR produces the putative conclusion *Some A are C* as its response (step 7).

Consider the initial encoding in more detail (step 1). VR begins by encoding the first premise *Some B are A* into an annotated model. Following the default encoding for *Some x are y* in Figure 3-3, it first looks for model objects with specified characteristics (e.g., having properties B and A). Since there are none (the annotated model is empty at this point), VR creates two new objects — (B') and (B' A)[8]. VR then encodes the second premise *All B are C*. Following Figure 3-3, it marks property B as identifying in all objects (a null operation in this case) and augments all objects having property B with property C. In this example, both objects are augmented. The result is an annotated model with two model objects (B' C) and (B' A C). This annotated model satisfies the generation templates for *Some B are A* and *All B are C* (Figure 3-4) and so these conclusions are proposed. Since neither is a legal syllogism conclusion (they do not relate the end terms), VR must resort to reencoding. The most recently modified model object is the recently augmented (B' A C) and C is its most recent property, so VR will begin reencoding using C as the reference property. Property A was the next most recent and so it will be tried next followed by B.

Steps 2 and 3 illustrate VR's reencoding when the reference property is C. Reencoding *Some B are A* extracts nothing about C (step 2). There is no way it could, since the premise does not even mention that property. The annotated model therefore remains unchanged and the same (illegal) conclusions are proposed. *All B are C* does mention property C and so there is a possibility that reencoding will extract new indirect information from this premise. In this example, however, VR does not realize it can extract anything about C from the premise and so once again the model is unchanged and generation fails (step 3).

[8] Throughout this example, the model objects in the figure will be ordered from the least recently accessed (at the top) to the most recently accessed (at the bottom). In this case, the two objects were created at the same time so the ordering is arbitrary. In practice, VR uses the ordering given in the figure.

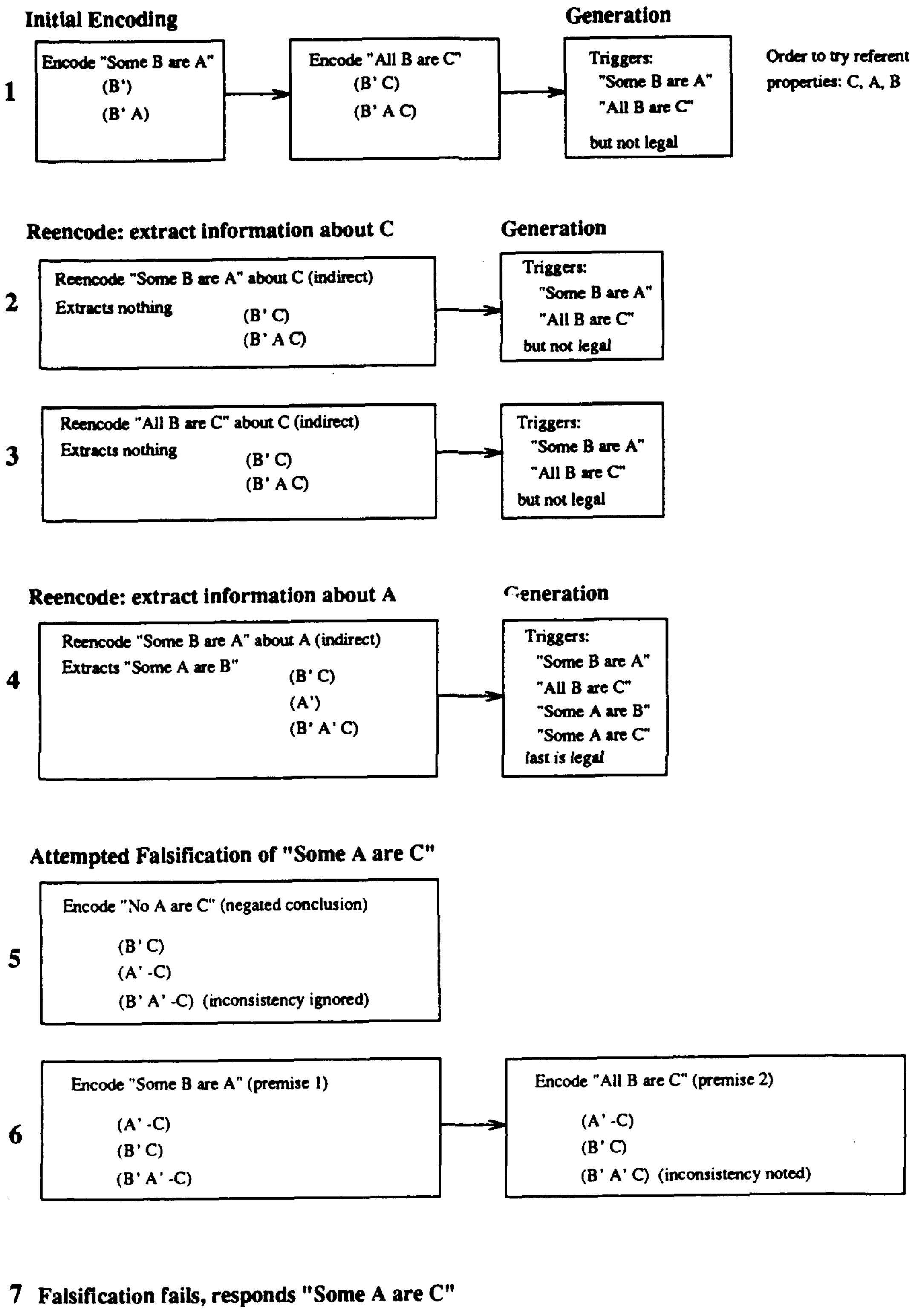

Initial Encoding
Generation
1
Encode "Some B are A"
(B')
(B' A)
Encode "All B are C"
(B' C)
(B' A C)
Triggers:
"Some B are A"
"All B are C"
but not legal
Order to try referent properties: C, A, B
Reencode: extract information about C
Generation
2
Reencode "Some B are A" about C (indirect)
Extracts nothing
(B' C)
(B' A C)
Triggers:
"Some B are A"
"All B are C"
but not legal
3
Reencode "All B are C" about C (indirect)
Extracts nothing
(B' C)
(B' A C)
Triggers:
"Some B are A"
"All B are C"
but not legal
Reencode: extract information about A
Generation
4
Reencode "Some B are A" about A (indirect)
Extracts "Some A are B"
(B' C)
(A')
(B' A' C)
Triggers:
"Some B are A"
"All B are C"
"Some A are B"
"Some A are C"
last is legal
Attempted Falsification of "Some A are C"
5
Encode "No A are C" (negated conclusion)
(B' C)
(A' -C)
(B' A' -C) (inconsistency ignored)
6
Encode "Some B are A" (premise 1)
(A' -C)
(B' C)
(B' A' -C)
Encode "All B are C" (premise 2)
(A' -C)
(B' C)
(B' A' C) (inconsistency noted)
7 Falsification fails, responds "Some A are C"

Figure 3-6: VR on *Some B are A, All B are C.*

VR then turns to reencoding the premises using reference property A (step 4). Unlike the previous attempts, reencoding *Some B are A* does extract some indirect knowledge. As mentioned earlier, this version of VR can extract the indirect knowledge *Some y are x* when reencoding *Some x are y* with reference property y. In this context, VR extracts the knowledge *Some A are B*. More precisely, it augments the annotated model in the same way it would when encoding that proposition. Following the specifications for *Some A are B* in Figure 3-3, it marks property A as identifying and creates a new object (A'). Since A has now become an identifying property, generation finally proposes conclusions about A in addition to those about B. One of these new conclusions (*Some A are C*) is legal and so generation succeeds and reencoding stops.

Unlike the default, this version of VR then attempts to falsify the putative conclusion (steps 5-6). It tries to construct an annotated model in which the putative conclusion is false, but the premises are true. To do so, it first encodes *No A are C* — the negation of *Some A are C* — so that the annotated model will *not* support the putative conclusion (step 5). In addition to adding -C to the middle model object, this encoding replaces property C in the bottom object with property -C. This change is flagged as an inconsistency, but is ignored since VR was expecting it. VR was intentionally encoding the negation of a proposition that was true in its annotated model. Doing so will usually lead to an inconsistency, but that does not necessarily mean that the negated conclusion is inconsistent with the premises (which is what VR is trying to determine). It is only if subsequently encoding the premises leads to an inconsistency that VR considers the falsification attempt to be a failure. So VR continues and encodes the premise *Some B are A*. The only change this causes to the model is to make the object (B' C) more recent than (A' -C) — it does not lead to an inconsistency. Encoding *All B are C*, however, causes property -C in the bottom object to be replaced with C and this inconsistency is noted. Consequently, VR considers the falsification attempt a failure (it was unable to produce an annotate. model in which the premises are true, but not the putative conclusion) and concludes that the putative conclusion is correct (step 7).

Chapter 4
Other Task Variants

So far we have only discussed generating a conclusion given a pair of premises, but numerous other variants of the task exist and VR should apply equally well to these. In this section, we will consider five standard task variants (determining the validity of a single conclusion, determining the validity of multiple conclusions, choosing the valid conclusion from a set of alternatives, solving syllogisms involving believed/disbelieved materials, and solving syllogisms whose premises have been elaborated to be unambiguous) and show how VR can be applied to each.

4.1. Testing a single conclusion

In one variant of the task (Janis & Frick, 1943), subjects are asked to determine the validity of a single given conclusion based on a pair of premises. There are at least two ways VR could be adapted to this version of the task. First, it could use a control structure almost exactly like that in Figures 3-1 and 3-5, but insist that generation produce the very conclusion being evaluated. This strategy is presented in Figure 4-1. If it succeeds in generating that conclusion (and if it also passes any attempts at falsification that the system attempts), then VR should consider it valid. On the other hand, if it gives up trying to generate the conclusion after repeated reencodings have failed, then all VR knows is that it was not able to generate it — not that the conclusion is invalid. Nevertheless, given the forced choice between valid and invalid in this situation, most subjects presumably choose invalid. After all, it was an attempt to *generate* the conclusion that failed — not an attempt to falsify it. Furthermore, the task is to determine whether the conclusion *necessarily* follows from the premises, suggesting that the subject should respond "invalid" if he has serious doubts. We will refer to this as the *generate-and-match strategy*.

A second approach, presented in Figure 4-2, is to abandon generation entirely. VR could just encode the premises followed by the conclusion and base its response on metaknowledge. If the conclusion is found to be inconsistent with the annotated model of the premises, then VR should respond that the conclusion is invalid. If it is redundant (and any attempts to falsify it fail), then VR should respond that the conclusion is valid.

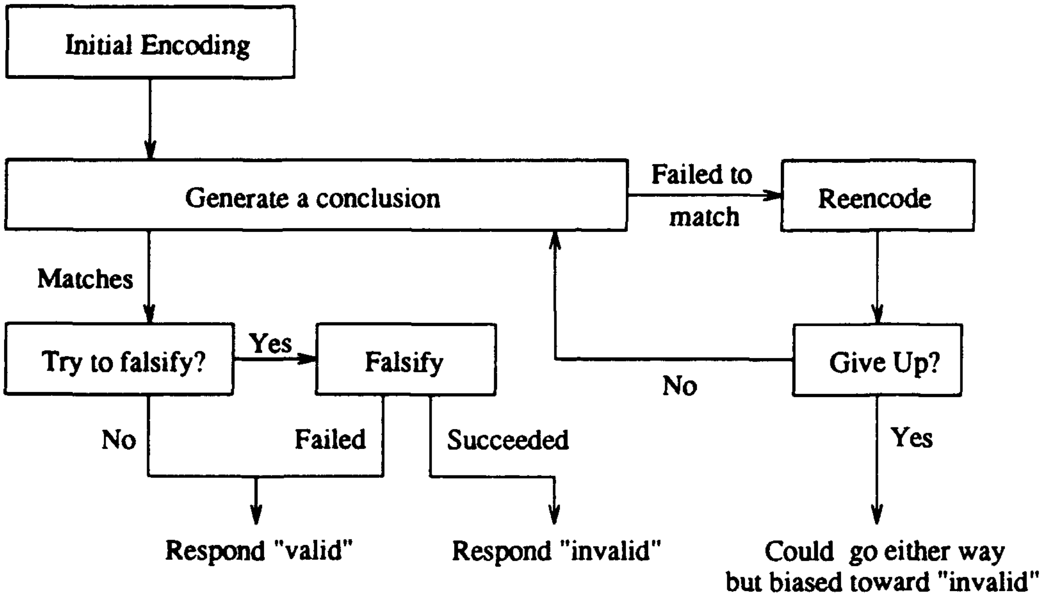


Figure 4-1: VR's generate-and-match strategy for testing conclusions.

If neither of these outcomes occurs, then VR does not know whether the conclusion is valid or not and should repeatedly reencode until it can decide or until it gives up. If it gives up, then it has no basis for making a decision and could go either way. Again, we assume most subjects would choose "invalid" under these circumstances since the task requires that valid conclusions *necessarily* follow from the premises. We will call this the *metaknowledge strategy*. These strategies are not mutually exclusive — the same subject could use both even within the same problem. For example, if one strategy failed to produce a definitive result, then trying the other would make sense.

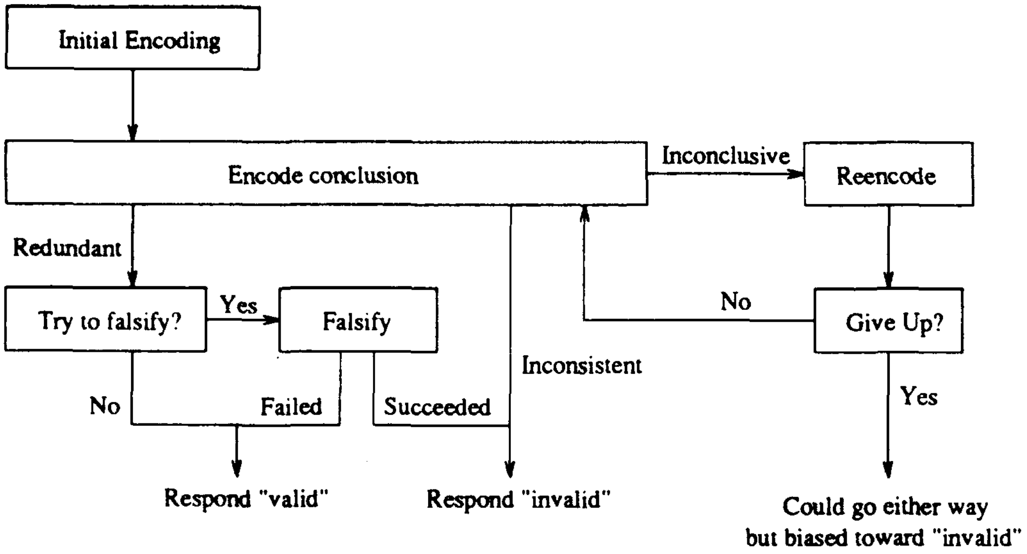


Figure 4-2: VR's metaknowledge strategy for testing conclusions.

4.2. Testing multiple conclusions

The two strategies above can be applied repeatedly to determine the validity of each member of a set of conclusions. This version of the task was used by Wilkins (1928) and Sells (1936) in their classic studies. Of course, it is possible that determining the validity of the first *k* conclusions could influence the response to the *k*+1st (e.g., there may be a bias to mark at least one of the alternatives as valid). If so, VR would require additional assumptions to account for such behavior.

4.3. Multiple choice

The most common variant of the task involves choosing the valid conclusion from a set of alternatives (Chapman & Chapman, 1959; Ceraso & Provitera, 1971, Revlis, 1975b; Dickstein, 1975). For this version of the task, a modified form of the generate-and-match strategy in Figure 4-1 is more efficient than the metaknowledge strategy. Of course, rather than responding "valid" or "invalid" the system should respond with the conclusion it has chosen. Also, if VR succeeds in falsifying a putative conclusion, it should go back to generation (like the original system in Figures 3-1 and 3-5). Although the generate-and-match strategy makes more sense, the metaknowledge strategy could still be used by repeatedly applying it to each alternative until one is found that is valid. If NVC is one of the options and none of the others can be validated, then it should be selected. If NVC is not an option, but none of the others can be validated, then VR should choose among those alternatives that were not determined to be invalid.

4.4. Believed/disbelieved conclusions

Numerous studies have asked subjects to solve syllogisms in which a conclusion contradicts or is confirmed by their existing beliefs (Wilkins, 1928; Janis & Frick, 1943; Morgan & Morton, 1944; Evans, Barston & Pollard, 1983; Oakhill & Johnson-Laird, 1985; Oakhill, Johnson-Laird & Garnham, 1989). VR suggests that subjects are often not certain that their reasoning is correct (e.g., they repeatedly reencode the premises in case they failed to extract some critical information) so that their beliefs cou! l easily influence their thinking. Exactly how beliefs influence VR's behavior depends on the specific version of the task it is working on.

If the task is to generate a conclusion from a pair of premises and VR produces a putative conclusion that contradicts beliefs (Oakhill & Johnson-Laird, 1985; Oakhill, Johnson-Laird & Garnham, 1989), then an obvious response is to go back and try to generate a different conclusion. If that fails (i.e., VR gives up before it succeeds), then VR must choose between its putative conclusion (which contradicted beliefs) and NVC. We assume different subjects make different choices under these conditions. Using this strategy, the believability of the conclusion has two effects on VR's processing: (1) it affects the criteria under which conclusion generation is considered successful (the "Succeeded" test for conclusion generation in Figures 3-1 and 3-5) and (2) it affects the response after giving up ("Respond NVC" in the same figures).

If the task is to determine the validity of one or more conclusions (Wilkins, 1928; Janis & Frick, 1943; Evans, Barston & Pollard, 1983), then the main effect shows up in how VR responds if the results of reasoning are inconclusive. Specifically, if VR uses the generate-and-match strategy, but is unable to generate a conclusion it believes to be true, then VR's decision would be biased by beliefs. Since encoding and reencoding are not guaranteed to be complete, valid conclusions can be missed. If a given conclusion is believed to be true, but VR cannot generate it, then it is natural for VR to assume that the encoding was faulty rather than the conclusion itself. Referring to Figure 4-1, the main effect of belief is on VR's response after giving up (bottom right of the figure).

Similarly, if VR tries to determine the validity of a given conclusion based on whether that conclusion is inconsistent or redundant with the annotated model of the premises (the metaknowledge strategy), and if the results of initial encoding and reencoding are inconclusive (i.e., the conclusion is not known to be inconsistent or redundant), we assume VR's decision is biased by its beliefs. Since its reasoning attempts failed to lead to a decision, VR may resort to basing its decision on what it believes to be true. Once again, belief has its main effect on VR's response to giving up (Figure 4-2, bottom right).

Even if the results of reasoning are conclusive, beliefs could still have an effect, though presumably a smaller one. For example, suppose the generate-and-match strategy succeeds in producing a disbelieved conclusion. VR could still assume that its reasoning attempts were faulty (they often are after all) and decide to base its decision on the believability of the conclusion (here belief affects the choice to "Respond 'valid'" in the bottom left of Figure 4-1). Similarly, if the metaknowledge strategy accepts or rejects a conclusion because it is redundant or inconsistent with the annotated model, but this choice is inconsistent with beliefs about the conclusion, then beliefs could influence the response ("Respond 'valid'" and "Respond 'invalid'" in Figure 4-2). Whereas the results of reasoning were inconclusive in the previous cases, here the results are definitive. As a result, the effects of belief are predicted to be smaller.

Finally, belief effects like those described above will also show up when VR tries to choose the valid conclusion from a set of alternatives (Morgan & Morton, 1944). Using the generate-and-match strategy, beliefs mainly affect the criteria under which generation is considered successful and how VR responds after giving up. VR's behavior is similar to that described above on the pure generation task — if it generates a conclusion that it does not believe then it will go back and try to generate a different one. If that fails, then it must make a choice between its putative conclusion (which it does not believe) and an alternative that does not contradict beliefs but that it is unable to generate. With the metaknowledge strategy, on the other hand, beliefs mainly influence how VR responds when the results of reasoning are inconclusive — just like they did when determining the validity of one or more given conclusions. If after encoding and reencoding the premises the status of the conclusion is inconclusive, then VR's response would be biased by its beliefs.

4.5. Disambiguated premises

In traditional syllogisms, . premise can often refer to many different states of affairs. For example, the premise *Some A are B* is consistent with situations in which all A are B and with situations in which some A are not B. To investigate whether this ambiguity influences performance, Ceraso & Provitera (1971) elaborated the premises so that they were unambiguous. For example, they used *Some A are B, Some A are not B, and Some B are not A* instead of *Some A are B* and *All A are B but Some B are not A* instead of *All A are B*. Their procedure corresponds to using more than two premises. Indeed, we can model this version of the task without any change to VR by explicitly presenting VR with the entire set of propositions that are presented to subjects in this experiment.

Chapter 5
Aggregate Data

5.1. Aggregate data

Figure 5-1 presents aggregate data from six experiments and results using three default versions of VR as well as their average. At the far right are the values that would be expected in purely random data. The top row in the figure presents the percentage of legal responses that were correct[9]. The other three rows present measures of the atmosphere, conversion and figural effects. Since the atmosphere effect does not apply to NVC responses, these were not included. Instead, the numbers represent the percentage of legal responses *other than NVC* that followed the atmosphere effect. The conversion effect also does not apply to NVC responses and, in addition, it is restricted to errors. Consequently, the numbers in the third row represent the number of legal but erroneous non-NVC responses that would be valid if the converse of one or both premises were true. Finally, the figural effect is only relevant to 32 of the 64 tasks (those in the xy-yz and yx-zy figures) and it too does not apply to NVC responses. The numbers in the bottom row therefore represent the percentage of legal non-NVC responses *on relevant tasks* that followed the figural effect.

	Unlimited	**Timed**	**Revised**	**Week 1**	**Week 2**	**Inder**	**Totals**	**VR1-3**	**VR1**	**VR2**	**VR3**	**Random**
	(n = 20)	(n = 20)	(n = 20)	(n = 20)	(n = 20)	(n = 3)	(n = 103)					
Correct (%)	40	49	45	59	69	61	53	58	64	75	38	15
Atmospheric (%)	69	72	76	81	84	84	77	89	100	100	83	25
Conversion (%)	28	38	38	38	51	54	37	40	50	29	41	9
Figural (%)	90	81	83	90	86	70	86	89	100	86	86	50

Figure 5-1: Percentage of correct, atmospheric, conversion and figural responses from humans, VR, and random data.

The first three data sets were collected from students at the University of Milan (Johnson-Laird & Bara, 1984). In the first (unlimited), subjects were given an unlimited amount of time to write down a response. In the second (timed), subjects were asked to respond

[9] Out of 6592 total subject responses, 159 (2.4%) were either illegible or did not correspond to one of the nine legal conclusions (many involved the middle term) and these were not included.

within ten seconds. Afterwards, they were given the same tasks along with the responses they gave in the timed condition and were given one minute to revise their answers if desired. These data are recorded in the third column (revised). The next two data sets (week 1 and week 2) were collected a week apart from the same students at Teachers College, Columbia University. The last data set (Inder) was collected from three undergraduates at the University of Edinburgh (Inder, 1986; 1987). None of the subjects had training in logic.

The three columns labeled VR1, VR2, and VR3 present predictions from three versions of VR, and column VR1-3 presents their average. VR1 is the default described previously. This version of the system is very streamlined — it does not extract any indirect knowledge when reencoding and it does not attempt to falsify its putative conclusions. Because it only extracts direct knowledge, VR1 often fails to relate the end terms and produces a large number of NVC responses (52/64). As a result, the measures of the atmosphere, conversion and figural effects for VR1 may not be reliable since they are based on a small sample of non-NVC responses (12/64 — 19%). In order to overcome this problem, we built two other versions of VR that do extract indirect knowledge and that, consequently, produce fewer NVC responses. VR2 is identical to VR1 except that it can extract two pieces of (correct) indirect knowledge: (1) when the target proposition is *Some x are y* and the reference property is y, it extracts *Some y are x*, and (2) when the target proposition is *No x are y* and the reference property is y, it extracts *No y are x*. This version of VR produces significantly more non-NVC responses (25/64) than VR1. VR3 is an extreme case that *always* extracts indirect knowledge and therefore produces very few NVC responses (8/64). Figure 5-2 presents the indirect knowledge that VR3 extracts from all combinations of target propositions and reference properties. The specific indirect knowledge extracted was chosen to be consistent with the target proposition, to relate the same two terms, and to share the same quantity (universal or particular).

	Reference Property		
Target Proposition	Y	-X	-Y
All X are Y	All Y are X	No non-X are Y	No non-Y are X
Some X are Y	Some Y are X	Some non-X are Y	Some non-Y are X
No X are Y	No Y are X	No non-X are Y	All non-Y are X
Some X are not Y	Some Y are not X	Some non-X are not Y	Some non-Y are not X

Figure 5-2: Indirect knowledge extracted by VR3.

The numbers in the far right column of Figure 5-1 (random) correspond to percentages that would be expected by chance alone. There are a total of 576 legal task-response pairs (64 tasks × 9 legal responses each). Of these, 85 (15%) represent a correct response to the associated task. This number is higher than the total number of tasks (64) because some tasks have more than a single correct response. If all 64 NVC responses are removed that leaves 512 task-response pairs that do not involve NVC. Of these, 128 (25%) are consistent with the atmosphere effect. Similarly, there are 464 task-response pairs that both represent errors and do not involve NVC. 40 of these (9%) would be valid if the converse of one or both premises were true. Finally, the figural effect does not apply to 256 of the 512 non-NVC task-response pairs (those in the yx-yz and xy-zy figures). Of the remaining 256, 128 (50%) follow the figural effect.

Figures 5-3 and 5-4 present percentages of responses that are consistent and inconsistent with beliefs. The numbers in the columns labeled B(%) correspond to the percentage of total responses that are assumed to be believed by the subjects (independent of their logical status), while the numbers in the columns labeled D(%) represent the percentage of disbelieved responses (syllogisms with the same logical form were used in the two cases). In most cases, believability ratings were collected from an independent set of subjects. Figure 5-3 presents four experimental paradigms involving testing conclusions, while the four in Figure 5-4 involve generating conclusions. Wilkins (1928), Janis & Frick (1943), and Morgan & Morton (1944) have also claimed evidence for a belief bias, but these studies did not include the critical comparisons between believed and disbelieved conclusions for syllogisms with the same logical form and so they have not been included in the figures. Janis & Frick (1943) used different logical forms for the believable and unbelievable conclusions. Morgan & Morton (1944) used the same logical forms but instead of comparing believed and disbelieved conclusions, they compared believed/disbelieved conclusions with abstract conclusions (involving X, Y, and Z rather than concrete terms). Wilkins (1928) used the same logical forms and always used concrete syllogisms, but compared believed/disbelieved with neutral conclusions rather than disbelieved with believed.

		Humans		VR1-3 Avg		VR1		VR2		VR3	
		B(%)	D(%)	B(%)	D(%)	B(%)	D(%)	B(%)	D(%)	B(%)	D(%)
Evans et al. (1983)	Expt 1	92	50	67	33	50	0	75	50	75	50
	Expt 2	76	38	63	25	50	0	75	50	63	25
	Expt 3	79	30	65	29	50	0	75	50	69	38
Oakhill et al. (1989)	Expt 3	58	52	75	50	75	50	75	50	75	50

Figure 5-3: Percentage of accepted conclusions that were believed and disbelieved in humans and VR.

The Evans, Barston & Pollard (1983) studies manipulated the believability of given conclusions while controlling their logical validity. In the first experiment, 24 undergraduates at Plymouth Polytechnic received the following two syllogisms: (1) *No A are B. Some C are B. Therefore, some C are not A* (valid), and (2) *No A are B. Some C are B. Therefore, some A are not C* (invalid). The same two syllogisms were presented with different contents so that the conclusion was either believable or unbelievable for a total of four tasks. The second experiment was similar but involved 64 undergraduates and four syllogism forms rather than two. In addition to the two listed above, subjects received the following additional syllogisms: (3) *Some A are B. No C are B. Therefore, some A are not C* (valid), and (4) *Some A are B. No C are B. Therefore, some C are not A* (invalid). Again, the syllogisms were presented using materials in which the conclusions were both believable and unbelievable leading to a total of eight tasks. In the third experiment, 32 first-year psychology students were presented with the syllogisms above as well as four others: (5) *No B are A. Some B are C. Therefore, some C are not A* (valid), (6) *No B are A. Some B are C. Therefore, some A are not C* (invalid), (7) *Some B are A. No B are C. Therefore, some A are not C* (valid), and (8) *Some B are A. No B are C. Therefore, some C are not A* (invalid). Oakhill, Johnson-Laird & Garnham (1989) (experiment 3) used a similar procedure with 30 subjects and the following six logical forms: (1) *All A are B. No B are C. Therefore, no A are C* (valid), (2) *Some A are B. All B are C. Therefore, some A are C* (valid), (3) *All A are B. All C are B. Therefore, all A are C* (invalid), (4) *All A are B. Some B are C. Therefore, some A are C* (invalid), (5) *No B are A. All B are C. Therefore, no A are C* (invalid), and (6) *Some B are A. No B are C. Therefore, some A are not C* (valid).

The first two columns in Figure 5-3 present the human data from the four experiments just described. The other columns present the predictions of three versions of VR as well as their average. These systems are the same as those described previously, but have been modified to respond to the new experimental paradigm. In particular, they test rather than generate conclusions (using the metaknowledge strategy in Figure 4-2) and they incorporate the same assumptions about the believability of conclusions that were used in the studies. We simulated the belief effects by having VR choose randomly between "invalid" and the response suggested by belief whenever it gave up (i.e., encoding and reencoding were inconclusive). We did not incorporate any belief effects when the results of reasoning were conclusive.

The four experiments presented in Figure 5-4 investigated belief bias effects when *generating* a conclusion rather than determining the validity of one that has been given. In these studies, subjects were presented with premise pairs and the experimenters manipulated the believability of the conclusion "suggested" by the premises. The suggested conclusions were selected based on Johnson-Laird's theories of syllogistic reasoning (Johnson-Laird, 1983; Johnson-Laird & Bara, 1984) — they corresponded to conclusions that are true in one mental model of the premises (the specific mental model that Johnson-Laird & Bara (1984) assumed to be built first) — and were confirmed to be common responses in the data he collected. In the first experiment (Oakhill & Johnson-Laird, 1985, experiment 1), the following two premise pairs were used with 24 subjects: (1) *Some A are not B. All C are B.* and (2) *Some A are not B. All B are C.*. In both cases, the suggested conclusion was assumed to be *Some A are not C.* These syllogisms were presented with different sets of materials so that the suggested conclusion would be either believable or unbelievable (they also manipulated whether the conclusion was empirically true/false (as a matter of fact) or definitionally true/false (by definition), but we have collapsed across these conditions). Figure 5-4 shows the percentage of generated conclusions that were believable (46%) compared with the percentage that were unbelievable (30%). Experiment 2 was similar but involved 16 subjects and the following two syllogisms: (1) *No A are B. Some C are B.* (suggested conclusion: *Some C are not A*) and (2) *Some A are B. No C are B.* (suggested conclusion: *Some A are not C*). Again, the syllogisms were presented with different materials so that the suggested conclusions were either believable or unbelievable. The other two experiments (Oakhill, Johnson-Laird & Garnham, 1989, experiments 1 and 2) used a similar procedure, but presented the following six premise pairs: (1) *All A are B. No B are C.* (suggested conclusion: *No A are C*), (2) *Some A are B. All B are C.* (suggested conclusion: *Some A are C*), (3) *All A are B. All C are B.* (suggested conclusion: *All A are C*), (4) *All A are B. Some B are C.* (suggested conclusion: *Some A are C*), (5) *No B are A. All B are C.* (suggested conclusion: *No A are C*), and (6) *Some B are A. No B are C.* (suggested conclusion: *No A are C*). In experiment 2, the indeterminate premise pairs (3 & 4) were presented twice as often as the others. 41 undergraduates participated in the first experiment and 21 in the second.

		Humans		VR1-3 Avg		VR1		VR2		VR3	
		B(%)	D(%)	B(%)	D(%)	B(%)	D(%)	B(%)	D(%)	B(%)	D(%)
Oakhill et al. (1985)	Expt 1	46	30	0	0	0	0	0	0	0	0
	Expt 2	49	43	50	25	0	0	100	50	50	25
Oakhill et al. (1989)	Expt 1	66	37	50	17	50	25	50	17	50	8
	Expt 2	74	28	50	15	50	25	50	13	50	6

Figure 5-4: Percentage of generated conclusions that were believed and disbelieved in humans and VR.

The first two columns in Figure 5-4 present the human data from these four experiments. The other columns present the predictions from the three versions of VR and their average. Once again, the systems are the same as before, but have been adapted to the demands of the task — they generate conclusions just like the original versions, but they incorporate a belief bias. Specifically, whenever the systems generate a conclusion that they do not believe, they go back and try to generate a different conclusion. If that fails then they choose randomly between their original conclusion and NVC.

Figure 5-5 presents data relevant to the elaboration effect. The top row presents the percentage of responses that were correct for a set of 13 standard syllogisms used by Ceraso & Provitera (1971). The bottom row presents the percentage of correct responses for the same syllogisms when their premises were elaborated to be unambiguous. The elaborated and unelaborated forms of the premises are shown in Figure 5-6. In the unelaborated condition, only the premises in boldface were presented while in the elaborated condition the others were included as well. Each syllogism was presented with four alternatives: (1) *All A are C*, (2) *Some A are C*, (3) *No A are C*, and (4) *Can't say*. The subjects were asked to choose the valid response from among these four.

	Humans	**VR1-3 Avg**	**VR1**	**VR2**	**VR3**
Unelaborated	58	59	62	69	46
Elaborated	80	95	100	100	85

Figure 5-5: Percent correct for elaborated and unelaborated premises in humans and VR.

1	2	3	4	5
All A are B All B are A **All B are C** All C are B	**All A are B** All B are A **All B are C** Some C are not B	**All A are B** All B are A **All C are B** Some B are not C	**All A are B** All B are A **Some B are C** Some B are not C Some C are not B	**All A are B** All B are A **No B are C**
6	**7**	**8**	**9**	**10**
All A are B Some B are not A **All B are C** Some C are not B	**All A are B** Some B are not A **All C are B** Some B are not C	**All A are B** Some B are not A **Some B are C** Some B are not C Some C are not B	**All A are B** Some B are not A **No B are C**	**All B are A** Some A are not B **All B are C** Some C are not B
11	**12**	**13**		
All B are A Some A are not B **Some B are C** Some B are not C Some C are not B	**Some A are B** Some A are not B Some B are not A **Some B are C** Some B are not C Some C are not B	**No A are B** **No B are C**		

Figure 5-6: Elaborated and unelaborated (bold) premises.

The human data (first column in Figure 5-5) was collected from 80 students at Rutgers-Newark (Ceraso & Provitera, 1971). The other columns present predictions from the three default versions of VR and their average. For this analysis, the VR systems were unchanged. The only difference between the conditions was whether the systems were given only the unelaborated premises or the entire set of elaborated premises (Figure 5-6).

5.2. Discussion

The human data in Figures 5-1, 5-3, 5-4 and 5-5 reflect all seven of the regularities from Figure 2-2. The first five can be seen in Figure 5-1. The difficulty of the task is reflected in the low percentage of correct responses (53% on average). Nevertheless, subjects performed far better than would be expected by chance alone (15% in the column on the far right) demonstrating a validity effect. In keeping with the atmosphere effect, 77% of the legal non-NVC responses were consistent with atmosphere compared with the 25% that would be expected by chance. A conversion effect is also apparent since 37% of the legal, erroneous non-NVC responses would have been valid with conversion compared with only 9% in random data. Finally, the data also reflect a strong figural effect — the percentage of relevant, legal, non-NVC responses that followed the figure of the premises in the human data (86%) is much higher than what would be expected in random data (50%). Notice that these five regularities are reflected in all six data sets indicating their robustness.

Figures 5-3 and 5-4 demonstrate a belief bias in both testing and generating conclusions. In all four experiments in Figure 5-3, the percentage of accepted conclusions that were believable was consistently larger than the percentage that were unbelievable (92% vs. 50%, 76% vs. 38%, 79% vs. 30% and 58% vs. 52%). This same effect can be seen in the four experiments that required generating rather than testing conclusions (Figure 5-4).

Finally, Figure 5-5 illustrates the elaboration effect. The percentage of responses that were correct rose significantly when the premises were elaborated to be unambiguous (58% vs. 80%).

These figures also demonstrate that all three default versions of VR produce these seven regularities. Looking at Figure 5-1, they only get between 38% and 75% of the syllogisms correct (difficulty effect), but this is much better than would be expected at random (validity effect). Also, far more of their responses follow the atmosphere (83-100% of non-NVC responses), conversion (29-50% of non-NVC errors) and figural effects (86-100% of relevant non-NVC responses) than would be expected in random data (25%, 9% and 50% respectively). Furthermore, Figures 5-3 and 5-4 show that all three systems accept and generate more believed than disbelieved conclusions (belief bias). In the one experiment in which the systems failed to show an effect of beliefs (Oakhill & Johnson-Laird, 1985, experiment 1), only two premise pairs were used. In both cases, VR1 & VR2 produced NVC while VR3 produced *Some C are not A*. These responses did not correspond to the "suggested" conclusion *Some A are not C* and, as a result, VR did not consider them either believable or unbelievable. Consequently, there was no opportunity for an effect of belief to show up. Similarly, the second experiment only involved two syllogisms and VR1 produced NVC on both so it did not produce a belief bias. In all other cases (and in all cases involving testing given conclusions), however, all three systems did produce a belief bias effect. Finally, Figure 5-5 demonstrates that the VR systems also exhibit an elaboration effect. All three systems produced a higher percentage of correct responses when the premises were elaborated to be unambiguous (between 85% and 100%) than when they were left unelaborated (46% to 69%).

These versions of VR not only produce the regularities, but the qualitative size of the effects is similar to that in the human data. The mean percent correct in the human data sets from Figure 5-1 ranges from 40% to 69% with a mean of 53% while for the three VR systems it ranges from 38% to 75% with a mean of 58%. Similarly, the size of the atmosphere effect in these data sets (ranges from 69% to 84% with a mean of 77%) is similar to that produced by the VR systems (83% to 100% with a mean of 89%) as is the size of the conversion (28% to 54% with mean 37% in human data — 29% to 50% with mean 40% for the VR systems) and figural effects (70% to 90% with mean 86% in human data — 86% to 100% with mean 89% for the VR systems). The quantitative predictions are less accurate in the case of belief bias (Figures 5-3 and 5-4), but aside from the top row in Figure 5-4 (which we discussed above), only two out of 14 absolute predictions are off by more than 20% and the predicted *size* of the effect is never off by that much. Finally, the percentage of elaborated and unelaborated syllogisms that the three VR systems solved correctly is also quite similar to the human data (59% vs. 58% for unelaborated premises and 95% vs. 80% for elaborated premises).

While these quantitative comparisons are interesting, the main point of these aggregate data is that a set of simple but different VR systems all produce the seven major regularities. Indeed, given that these systems were constructed as simple defaults (rather than as models of a "modal" or "typical" subject), it is actually surprising that their quantitative predictions are as accurate as they are. The important point is that a variety of instantiations of VR all produce the seven major phenomena of syllogistic reasoning.

5.3. Explanations of the regularities

5.3.1. Difficulty effect

Since VR predicts these regularities, analyzing how it does so will suggest explanations for why they occur. First consider the question of why syllogisms are so difficult. VR suggests that the main reason is that language comprehension is not guaranteed to deliver a necessary and sufficient representation of what the premises are about. It just constructs a model of a situation that is *consistent* with the premises (i.e., the premises will be true in the model). Such a model can both support conclusions that are not strictly valid (and be unwarranted) and fail to support conclusions that are (and be incomplete). Figure 3-2 gave examples of each.

5.3.2. Validity effect

The assumption that annotated models are always consistent, if not always necessary and sufficient, also provides an explanation of the validity effect — the fact the people perform better than chance. For an annotated model to be consistent with the premises implies that no entailments of the premises can be false in the model (this does not mean that the entailments must be true in the model). Otherwise one (or both) of the premises would have to be false. Equivalently, a consistent annotated model never supports conclusions that contradict an entailment of the premises. So if people base their conclusions on an annotated model that is consistent with the premises, their responses will always be possible (though not necessary) — they will not consider (invalid) conclusions that contradict an entailment of the premises. Consequently, they do better than just randomly selecting from all possible conclusions.

5.3.3. Atmosphere effect

The reason that VR produces the atmosphere effect and, hence, the suggested explanation, is that the standard semantics of the premises rule out most non-atmospheric conclusions. Assuming that "some" is typically interpreted as "some but not necessarily all" (or just "some but not all") explains why particular responses are so common for *syllogisms that involve* a particular premise (but not for those that don't). When a "some" premise is read it creates additional model objects that do not satisfy the predicate of the premise (the lower conditional sequences for "Some" and "Some not" in Figure 3-3 created such objects). These additional model objects rule out universal conclusions and lead to particular responses. For example, the default encodings in Figure 3-3 lead to the following annotated model for the premise *Some A arc B*:

```
(A' )
(A'  B)
```

where the upper model object encodes the "not all" information. Augmenting this model based on *All B are C* leads to:

```
(A' )
(A'  B'  C)
```

and the upper model object refutes the universal conclusion *All A are C* leading to a "some" response instead. If both premises are universal, these additional objects are not created and universal conclusions can be drawn. If the interpretation of "some" during generation is similar to that during comprehension ("some but not (necessarily) all") then particular conclusions will only be drawn if universals cannot. Consequently, two universal premises tend to produce a universal conclusion.

In a similar way, negative premises tend to refute positive conclusions since they create negative rather than positivc associations between properties. For example, consider reading *All A are B* and *No B are C* using the default encodings from Figure 3-3. *All A are B* leads to the single model object (A' B) and then *No B are C* augments it with -C — (A' B' -C) — creating a negative rather than positive association. Since negative premises tend to produce negative associations, the resulting conclusions are also negative. Conversely, if both premises are positive then positive rather than negative associations among properties are created and conclusions tend to be positive.

5.3.4. Conversion effect

There are two main reasons VR produces a conversion effect. The first is straightforward — VR often extracts indirect knowledge that is equivalent to the converse of a premise (even if that converse is invalid). For example, when VR3 tries to extract knowledge about y from a premise that relates x to y, the knowledge it encodes is always equivalent to the converse of the premise (Figure 5-2, left column). In two of the four cases, the converse is invalid and can lead to errors consistent with the conversion effect. Consider the processing that VR3 goes through while working on the following premise pair: *Some A are B, All C are B*. Initial encoding leads to the following annotated model:

```
(A')
(A'  B)
(C'  B)
```

and B is chosen as the first reference property. Reencoding the first premise with respect to B only changes the recency of some of the objects, but reencoding *All C are B* with respect to B augments the model in the same way encoding *All B are C* would:

```
(A')
(C'  B')
(A'  B'  C)
```

which supports both *Some A are C* and *Some C are A*. Both responses are consistent with the conversion effect (i.e., they are invalid but would be correct if the converse of one or both premises were true).

This explanation is reminiscent of previous theories of syllogistic reasoning that assumed illicit conversion (Chapman & Chapman, 1959; Revlis, 1975b), but there are important differences. For one, those theories attempted to explain the majority of non-NVC errors in terms of illicit conversion while we assume there are a variety of other sources of error. Indeed, as shown in Figure 5-1 only 40% of non-NVC errors produced by VR1, VR2 and VR3 can be explained in terms of illicit conversion, but this simulates the human data very well (in which only 37% or non-NVC errors were consistent with the conversion effect). The fact that around 60% of non-NVC errors *cannot* be explained in terms of illicit conversion poses problems for the Chapman & Chapman (1959) and

Revlis (1975b) theories, but not for VR. Furthermore, Revlis intentionally implemented the strongest version of the conversion hypothesis, namely, that both premises are always explicitly converted and that both they and their converses are encoded at the same time (as Revlis points out, Chapman & Chapman were vague on this point). VR makes no such assumptions. In VR, indirect knowledge is only extracted if direct knowledge fails to lead to a conclusion. Even if indirect knowledge corresponding to illicit conversion is extracted, the converted premise may not be explicitly available[10].

VR1 and VR2 both produce conversion effects (Figure 5-1) and yet neither extracts invalid indirect knowledge (VR2 only extracts valid indirect knowledge and VR1 does not extract any at all). According to VR then, there must be at least one other factor that contributes to errors consistent with the conversion effect. The source of most of these other errors is the traditional *fallacy of the undistributed middle term* and shows up in the interpretation of particular premises. Consider the premise pair *All A are B, Some B are C*. Both VR1 and VR2 behave the same way on this problem. Encoding the first premise creates a single model object (A' B) and the second premise augments it as well as creating a new one:

```
(A' B)
(A' B' C)
```

This annotated model leads to the invalid conclusion *Some A are C*. Furthermore, this conclusion is consistent with the conversion effect since it would be valid if the converse of *All A are B* (that is, *All B are A*) were true. But no knowledge corresponding to the converse of either premise has been extracted (indeed, no indirect knowledge has been extracted at all). The problem is that *Some B* was interpreted as referring to a subset of the objects mentioned in the first premise, and this interpretation is invalid (to see this, consider the syllogism *All cats are animals, Some animals are dogs. Therefore, some cats are dogs*). So some errors consistent with the conversion effect can be explained in terms of the erroneous interpretation of particular premises, without any reference to illicit conversion.

5.3.5. Figural effect

VR's explanation of the figural effect is based on the special availability of proposition topics in the annotated model. After encoding, properties in the annotated model that correspond to premise topics are more available than other properties (they are marked as identifying properties). During generation, conclusions about those properties are tried

[10] The assumption in VR is that indirect knowledge augments the annotated model in the same way encoding a corresponding explicit proposition would. It does not assume that the indirect knowledge is available as an explicit proposition. It may or may not be — VR does not take a stand on the issue.

first. If only one end term appeared as the topic of a premise, then the corresponding property will be marked as identifying, and generation will tend to produce conclusions about it. In the xy-yz and yx-zy figures, this leads to xz and zx conclusions, respectively, as predicted by the figural effect.

5.3.6. Belief bias

At the most general level, there are two reasons VR is biased by beliefs. First, its reasoning attempts are often inconclusive so that it looks for other knowledge sources on which to base a decision. Beliefs are such a knowledge source. More specifically, beliefs can influence how VR responds after giving up (by responding based on beliefs rather than following a default). Second, since VR's reasoning attempts can be faulty, it will reconsider results it would normally accept as valid if they contradict beliefs. For example, beliefs can influence the criteria under which conclusion generation is considered successful (the "Succeeded" test for conclusion generation in Figures 3-1 and 3-5). If VR generates a conclusion that contradicts beliefs, then it will go back and try to generate a different one, even if that conclusion was legal.

5.3.7. Elaboration effect

VR's behavior on the premises in Figure 5-6 suggests that part of the elaboration effect on these tasks is due to artifacts in the experiment. Specifically, there are two main problems with the Ceraso & Provitera (1971) study that could account for part of the elaboration effect they observed: (1) it only presents four responses from which to choose and (2) the valid conclusions are sometimes different for the elaborated and unelaborated premises. Consider tasks 8 and 12 in Figure 5-6. The VR systems produce the erroneous response *Some A are C* to the unelaborated forms of both tasks, but give the correct response (NVC) in the elaborated forms. These systems do generate one conclusion (*Some C are not A*) while working on the elaborated premises, but since this conclusion is not one of the alternatives and since they cannot produce any others, they respond with NVC. In a different version of the task (that included this as a legal choice) these systems would also have gotten the elaborated tasks wrong. Task 4 (Figure 5-6) demonstrates a different problem with the experiment. In the unelaborated form, the correct response to this task is NVC, but in the elaborated form it is *Some A are C*. In both formats, all 3 VR systems respond *Some A are C* and they appear to exhibit an elaboration effect. The problem is that it is impossible to tell whether elaboration has had any effect. Even a theory that assumed elaborations would have no impact on performance could exhibit an elaboration effect on this task as long as it responded *Some A are C*.

An obvious question is whether there really is an elaboration effect at all, or whether it can be completely explained by these kinds of artifacts. Figure 5-7 presents data relevant to this question. These data are the same as those in Figure 5-5 except that the three tasks discussed above have not been included. Task 3 has also been excluded since the correct response changes between the unelaborated and elaborated forms (like task 4). Notice that although these tasks have been removed, both the human data and VR's predictions still show an elaboration effect (though it is smaller than before).

	Humans	VR1-3 Avg	VR1	VR2	VR3
Unelaborated	70	78	78	89	67
Elaborated	80	96	100	100	89

Figure 5-7: Percent correct for elaborated and unelaborated premises in humans and VR (with four tasks excluded).

VR suggests two explanations for why this effect continues to show up. First, elaborated premises can constrain the annotated model to represent objects that fail to relate the end terms (so that invalid universal conclusions are not generated). Premise elaborations often refer to objects that are not referred to in the premises themselves and that do not associate the end term properties. Since the annotated model is constrained to be *consistent* with the propositions it encodes (i.e., the propositions are true in the model), it *must represent* these objects explicitly when given elaborated premises. Since these objects do not relate the end terms, invalid universal conclusions (that would otherwise be generated) are not proposed.

A second reason VR produces an elaboration effect is that elaborations can make certain properties more available (by marking them as identifying), causing conclusions to be generated that might otherwise be overlooked. In some cases, VR will produce an annotated model that relates the end terms, but still be unable to generate a legal conclusion because the property that should be the topic does not appear as an identifying property. But if that property appears as the topic of a premise elaboration, then it will marked as identifying and the appropriate conclusion can be drawn.

Both these effects can be seen in VR1's behavior on task 10 in Figure 5-6. After initial encoding and reencoding (VR1 does none) of the unelaborated premises, *All B are A, All B are C*, VR1 constructs the following annotated model:

```
(B' A C)
```

Although this model relates properties A and C, VR1 cannot draw a conclusion because neither of these properties is marked as identifying. Consequently, it responds with NVC. Notice that even if property A *had* been an identifying property, however, VR1 would have responded with *All A are C* which is still incorrect (this is the response given by VR3 to this premise pair). In contrast, if the original premises had been elaborated to include the additional premises *Some A are not B* and *Some C are not B*, then the annotated model would have been:

```
(A')
(A'  -B)
(B'  A C)
(C')
(C'  -B)
```

The premise elaborations have caused two significant changes to the annotated model: (1) properties A and C now appear as identifying properties and (2) additional objects that do not relate the end terms have been included. As a result, VR1 produces the valid conclusions *Some A are C* and *Some C are A* (and only those conclusions). *Some A are C* is one of the four legal alternatives and so it is selected.

It is worth mentioning that there are at least two other factors that could also lead to elaboration effects, but that VR has not simulated. First, premise elaborations could improve performance by blocking the extraction of contradictory indirect knowledge. Second, elaborated premises could influence whether existing model objects are augmented or new objects are created. For example, VR would predict that using elaborated particular premises such as *Some* other *B are C* instead of just *Some B are C* could significantly improve performance by helping subjects to avoid the fallacy of the undistributed middle term (page 41). We have not simulated these types of effects, but they are consistent with the assumptions inherent in VR.

5.4. Other predictions

VR makes a number of other predictions about syllogistic reasoning that are not included in the list of standard regularities. We will now consider some of VR's major assumptions and derive predictions from each.

First, consider initial encoding. As we have stressed, we assume initial encoding delivers an annotated model that is consistent with the premises. This assumption predicts that erroneous non-NVC responses are much more likely to be consistent with the premises than would be expected by chance alone. Equivalently, non-NVC errors are less likely to contradict an entailment of the premises than would be expected at random. To test this prediction we analyzed all 2424 legal, non-NVC errors that occurred in the six data sets presented in Figure 5-1. Of these, only 14 (0.6%) contradicted an entailment of the premises. At random, one would expect 251 (10%) of these non-NVC errors to be inconsistent with the premises.

Initial encoding also assumes that direct knowledge (about a topic) is more available than indirect knowledge. Specifically, during initial encoding only direct knowledge is extracted and even during reencoding, indirect knowledge is not guaranteed to be available though direct knowledge is (e.g., VR1 is never able to extract indirect knowledge). This assumption leads to a number of predictions about the xy-zy figure. In this figure, the topic of one premise is not mentioned in the other premise. Since direct knowledge is always about the topic of a proposition, this implies that without any indirect knowledge, the annotated models in the xy-zy figure will never relate the end terms. Instead, the first premise will create objects with property X and the second will create new objects with property Z. It is only through indirect knowledge about Y that the end terms could be related. In contrast, indirect knowledge is usually not necessary to relate the end terms in the xy-yz and yx-zy figures (we will discuss the yx-yz figure below). This leads to a number of predictions. First, there will be more NVC responses in the xy-zy figure than in either the xy-yz or yx-zy figures. Since indirect knowledge is required in the xy-yz figure, but not the other two, the annotated model is less likely to relate the end terms and more NVC responses should be observed. Second, subjects should be more accurate on indeterminate premise pairs (those without a valid conclusion) in the xy-zy figure than on those in either the xy-yz or yx-zy figures. If NVC responses are more common, then performance should improve on tasks for which this is the correct response. Conversely, subjects should be less accurate on determinate premise pairs (those *with* a valid conclusion) in the xy-zy figure than in either the xy-yz or yx-zy figures.

All but one of these predictions are confirmed in the six data sets from Figure 5-1. NVC responses comprise 55% of legal responses in the xy-zy figure, compared with only 33% and 41% in the xy-yz and yx;zy figures respectively. 67% of legal responses to indeterminate premise pairs in the xy-zy figure are correct, while only 45% and 52% are correct in the xy-yz and yx-zy figures. Conversely, only 45% of legal responses to determinate premise pairs are correct in the xy-zy figure compared with 55% in the xy-yz figure. Contrary to our prediction, only 46% of legal responses to determinate premise pairs are correct in the yx-zy figure and this is comparable to the 45% in the xy-zy figure. There are two determinate tasks in the yx-zy figure that are much harder than any of the others (*All B are A, No C are B* which no subject solves correctly and *Some B are A, No C are B* which only 10% of subjects solve correctly). For both, the valid conclusion is *Some A are not C* which goes against the figural effect. It could be that the predicted effect is present, but that it is being obscured by the figural effect on these two tasks (recall that the figural effect does not apply to the xy-zy figure). Consistent with this view, if these two tasks are excluded from the analysis, the percent correct on determinate yx-zy tasks jumps to 67% (compared with 45% in the xy-zy figure). Even if the two hardest determinate tasks in the xy-zy figure are excluded, the percent correct only increases to 56% which is still significantly lower.

The assumptions underlying VR's conclusion generation make similar predictions in the yx-yz figure. Recall that identifying properties are tried first during generation. But in the yx-yz figure, neither end term appears as the topic of a premise so they will not be marked as identifying after initial encoding. In this figure, then, indirect knowledge is required for an end term to become identifying. Since this is not the case in the xy-yz and yx-zy figures, VR makes the same predictions about the relationship between yx-yz and xy-yz/yx-zy tasks as it does for the xy-zy figure: there will be more NVC responses, indeterminate tasks will be easier, and determinate tasks will be harder. All these predictions are confirmed in the human data from Figure 5-1. NVC responses comprise 55% of the legal responses in the yx-yz figure compared with 33% and 41% in the xy-yz and yx-zy figures, 80% of legal responses to indeterminate tasks in this figure are correct as opposed to 45% and 52% in the other figures, and only 36% of legal responses to determinate tasks are correct compared with 55% and 46% in the others.

Turning to reencoding, VR assumes that whenever the annotated model fails to lead to a legal conclusion, subjects reencode the premises in an effort to extract more knowledge and augment their model. If subjects are constrained to respond within a short period of time, then they may be forced to give up before they have completed all their reencoding attempts. Consequently, they may miss conclusions that they would produce with more time. This predicts a higher proportion of NVC responses from timed subjects compared with the same subjects given unlimited time (at least when generating conclusions rather than evaluating them). The data sets labeled "Timed" and "Revised" in Figure 5-1 provide a relevant comparison. In the timed experiment, subjects received all 64 premise pairs and were asked to respond within 10 seconds. Afterwards, they were presented with the same tasks along with the responses they had just given and were given one minute to revise their answers if desired (revised). Consistent with VR's prediction, NVC responses comprised 56% of legal responses in the timed condition compared with 47% in the revised condition. It is possible that providing subjects with their previous responses could account for this effect if they considered NVC responses to be more tentative than others and tried harder to revise them (in VR, NVC is a response to giving up and so it may be more tentative). This interpretation becomes less likely, however, when we consider the other data sets from Figure 5-1. In three of the four (unlimited, week 1 and week 2), the percentage of NVC responses was below the 56% observed in the timed data (34%, 45% and 48%). And while two of these studies used completely different subjects, the first (unlimited, in which only 34% of responses were NVC) used subjects from the same subject pool as the timed experiment (students at the University of Milan). The one study that did not show a smaller percentage of NVC responses (Inder with 56%), involved only 3 subjects. So while the evidence is not conclusive, the data seem to confirm VR's prediction that time constraints should lead to more NVC responses.

Figure 5-8 summarizes the above predictions along with their empirical status.

Prediction	Empirical Status
1. Non-NVC errors more likely to be consistent with premises than expected by chance.	Confirmed
2. More NVC responses in xy-zy figure than xy-yz.	Confirmed
3. More NVC responses in xy-zy figure than yx-zy.	Confirmed
4. More correct responses to indeterminate xy-zy tasks than indeterminate xy-yz tasks.	Confirmed
5. More correct responses to indeterminate xy-zy tasks than indeterminate yx-zy tasks.	Confirmed
6. Fewer correct responses to determinate xy-zy tasks than determinate xy-yz tasks.	Confirmed
7. Fewer correct responses to determinate xy-zy tasks than determinate yx-zy tasks.	Marginally confirmed
8. More NVC responses in yx-yz figure than xy-yz.	Confirmed
9. More NVC responses in yx-yz figure than yx-zy.	Confirmed
10. More correct responses to indeterminate yx-yz tasks than indeterminate yx-zy tasks.	Confirmed
11. More correct responses to indeterminate yx-yz tasks than indeterminate yx-zy tasks.	Confirmed
12. Fewer correct responses to determinate yx-yz tasks than determinate yx-zy tasks.	Confirmed
13. Fewer correct responses to determinate yx-yz tasks than determinate yx-zy tasks.	Confirmed
14. More NVC responses under time constraints.	Marginally confirmed

Figure 5-8: Other predictions from VR and their empirical status.

Chapter 6
Individual Data

Data from individual subjects provide an even more stringent test for a theory than do aggregate data. Nevertheless, essentially no work in the reasoning literature has attempted to analyze and predict such data. A computational system like VR, however, provides a natural approach to attacking this issue. The first step is to identify those aspects of VR that could most plausibly differ across subjects and to formulate these as a set of explicit parameters. Then, VR can be run with different parameter settings in order to tailor its performance to that of individual subjects.

6.1. Individual difference parameters for VR

Figure 6-1 presents such a set of individual difference parameters. The first six parameters affect how VR encodes propositions. The next three (7-9) influence VR's conclusion generation process. Parameters 10 through 21 control what indirect knowledge VR extracts during reencoding. Finally, parameter 22 controls whether or not VR attempts to falsify its putative conclusion. We also considered including a parameter to control when VR would give up, but the only plausible alternative value we came up with (give up immediately — do not do any reencoding at all) corresponds to setting all the indirect knowledge parameters to "a". So we decided not to include that parameter.

This parameter set represents what we believed to be the most plausible sources of individual differences in syllogisms, but it is not meant to be exhaustive — there could be other parameters and additional values for those that have been included. But, as we will see, analyzing this set can point out other sources that we initially overlooked as well as indicating whether or not each of the posited parameters is important.

We assume the premises *Some x are y* and *Some x are not y* are often interpreted to mean more than just "there exists an object that has properties X and Y (-Y)". Parameters 1 and 2 control what additional information is extracted. If they are set to "a", then VR also ensures that there are other objects with property X that do not have Y or -Y (they may or may not be Y). This interpretation corresponds to the default encoding given in Figure 3-3. With value "b" VR creates a new (different) object that only has property X

Compound semantics of premises

1. Some X are Y and ...
- + a. other x may or may not be y
- + b. different x may or may not be y
- c. different x are not y
- - d. other x are not y and other x may or may not be y
- - e. nothing

2. Some X are not Y and ...
- + a. other x may or may not be y
- + b. different x may or may not be y
- c. different x are y
- - d. other x are y and other x may or may not be y
- - e. nothing

Atomic semantics of premises

3. All X are Y
- + a. all (x) → (x' y)
- - b. all (x) → (x' y) and new (x' y)

4. Some X are Y
- - a. MR (x y) → (x' y) else MR (x [not -y]) → (x' y) else new (x' y)
- + b. MR (x [not -y]) → (x' y) else new (x' y)
- + c. MR (x y) → (x' y) else new (x' y)
- d. new (x' y)

5. No X are Y
- + a. all (x) → (x' -y)
- - b. all (x) → (x' -y) and new (x' -y)

6. Some X are not Y
- - a. MR (x -y) → (x' -y) else MR (x [not y]) → (x' -y) else new (x' -y)
- + b. MR (x [not y]) → (x' -y) else new (x' -y)
- + c. MR (x -y) → (x' -y) else new (x' -y)
- d. new (x' -y)

Generation templates

7. Some X are Y
- + a. (x' y) (x [not y])
- b. (x' y)

8. Some X are not Y
- + a. (x' -y) (x [not -y])
- b. (x' -y)

Topics to try

9. Generation topics
- + a. only identifying properties
- b. identifying and then secondary properties

Indirect knowledge extracted about Y from ...

10. All X are Y
- + a. none
- + b. All y are x
- c. Some y are not x
- - d. There exists a different y

11. Some X are Y
- + a. none
- + b. Some y are x

12. No X are Y
- + a. none
- + b. No y are x

13. Some X are not Y
- + a. none
- - b. Some y are x
- + c. Some y are not x

Indirect knowledge extracted about -X from ...

14. All X are Y
- + a. none
- b. No non-x are y
- - c. All non-x are y

15. Some X are Y
- + a. none
- b. Some non-x are not y
- c. Some non-x are y

16. No X are Y
- + a. none
- - b. All non-x are y
- c. No non-x are y

17. Some X are not Y
- + a. none
- - b. Some non-x are y
- c. Some non-x are not y

Indirect knowledge extracted about -Y from ...

18. All X are Y
- + a. none
- b. No non-y are x

19. Some X are Y
- + a. none
- b. Some non-y are not x
- c. Some non-y are x

20. No X are Y
- + a. none
- b. All non-y are x

21. Some X are not Y
- + a. none
- - b. Some non-y are x
- c. Some non-y are not x

22. How to falsify
- + a. don't
- b. do, if succeed then NVC

Figure 6-1: Individual difference parameters for VR.

(it may or may not be Y)[11]. Value "c" is similar except that the new object also has property -Y. Value "d" leads to two other objects with property X — one with -Y and one that may or may not be Y. Finally, with value "e", no additional information is extracted at all. We assume that the compound semantics of the universal premises (*All* and *No*) are much less ambiguous and so we do not have such parameters for these premises.

The next four parameters (3-6) specify the atomic semantics of the four premise types independent of any additional information that may be extracted. For the universal premises (parameters 3 and 5), both interpretations augment all existing X's with property Y (or -Y), but with value "b" they also create a new object (this additional object encodes the fact that there may be additional X's that have not yet appeared in the annotated model). For the particular premises (parameters 4 and 6), the alternative interpretations differ in what X's they refer to (i.e., in what model object ends up getting augmented). Consider parameter 4. With value "a", the most recently accessed (MR) object with properties X and Y is chosen and X is marked as identifying. If no such object exists, then the most recently accessed object with X but not with -Y is augmented. If, again, there is no such object then a new object (X' Y) is created. This interpretation corresponds to the default in Figure 3-3. Interpretation "b" is similar, but it immediately looks for an X that lacks -Y while "c" gives up (and creates a new object) if it cannot find an object with both X and Y. Value "d" always creates a new object.

Parameters 7 and 8 control the generation templates for *Some x are y* and *Some x are not y*. Value "a" for parameter 7 corresponds to "some but not (necessarily) all x are y" — in addition to an X that is a Y, there must be an X which is not known to be a Y (it could be a -Y, but does not have to be). Value "b" corresponds to "some and possibly all" — even if all X's are Y's the particular conclusion will be proposed. Parameter 9 controls whether generation can produce conclusions about secondary (non-identifying) properties. With value "a", generation never produces conclusions about secondary properties. With value "b", secondary properties will be tried, but only after identifying properties.

Parameters 10 through 21 control what indirect knowledge is extracted during

[11] The distinction between "different" and "other" objects is subtle but important. It basically corresponds to whether the additional information is assumed to have the same referent as the primary information (e.g., whether or not *Some x may or may not be y* refers to the same set of objects with property X as does *Some x are y*). If they have the same referent ("other" rather than "different"), then any other properties that the referent has (e.g., Z) will appear in both objects. If they do not have the same referent ("different"), then the additional information will create a completely new object with at most two properties (X and Y or -Y). There may be cases in which the primary referent actually contradicts the additional information (e.g., the referent (X Y) contradicts *some x are not y*). Under these circumstances, even interpretations corresponding to "other" will create new objects.

reencoding. The first four (10-13) specify what is extracted when the reference property is y and the premise being reencoding relates X to Y. The other two sets of four reencoding parameters specify what is extracted about -X (14-17) and -Y (18-21).

Finally, parameter 22 controls falsification. With value "a", VR does not try to falsify its putative conclusions. With value "b" it does try to falsify and if it succeeds then it responds with NVC.

6.2. Fitting VR to individual data

As previously mentioned, while the parameter space in Figure 6-1 is almost certainly not exhaustive, it was designed to capture all of the sources of individual differences that we believe to be most important. Consequently, it is quite large (~3.6 billion parameter settings). In order to substantially reduce the size of this space, we performed some preliminary analyses in order to identify parameter values whose removal would least affect VR's fit to the individual data.

We divided the 64 premise pairs into four sets of 16 tasks such that each set contained the same number of tasks from each figure and of each premise type[12] and randomly selected 20 subjects from the 103 in Figure 5-1[13]. We then computed the best-fit parameter settings for each subject on all four task subsets — that is, the parameter settings that caused VR to produce the most correct predictions on those tasks.

It may seem surprising that we were able to find all the optimal parameter settings for each subject in such a large parameter space (~3.6 billion parameter settings) within a reasonable amount of time. Indeed, this was a major problem and led us to develop an entire suite of computational tools (collectively called ASPM — Analysis of Symbolic Parameter Models) for the analysis and fitting of parameterized models like VR (though ASPM has also been applied in other domains). We used ASPM to perform all the individual difference analyses described here as well as many others. For a complete description of ASPM and examples of its use see Polk, Newell, & VanLehn (1992).

[12] The specific tasks in each set were (using the notation from the middle of Figure 2-1): (1) {ObaEcb, EbaEcb, AbaOcb, AbaIcb, ObaEbc, ObaAbc, IbaObc, AbaEbc, EabIcb, IabAcb, AabEcb, AabAcb, OabObc, OabIbc, IabEbc, IabIbc}, (2) {AabIbc, EabOcb, EbaAbc, ObaAcb, IabOcb, IbaEbc, EabEbc, ObaOcb, IabIcb, AbaAbc, ObaIcb, OabAbc, AabIcb, EbaOcb, EbaIbc, EabAbc}, (3) {ObaIbc, OabAcb, AbaEcb, EabIbc, IbaAcb, EabObc, IbaIbc, EabEcb, ObaObc, AbaAcb, IabObc, OabEcb, IabAbc, EbaIcb, EabAcb, AbaObc}, and (4) {AabEbc, OabIcb, EbaObc, IbaEcb, IbaAbc, AabObc, OabOcb, IbaIcb, EbaEbc, AabAbc, IbaOcb, IabEcb, AbaIbc, OabEbc, AabOcb, EbaAcb}.

[13] Subjects 2 and 11 from unlimited, 4, 13, and 18 from timed, 3, 5, 6, 11, 15, and 16 from revised, 5, 7, 8, and 13 from week 1, 2, 6, 13, and 20 from week 2, and subject 1 from the Inder subjects.

Finally, we computed the frequency with which each parameter value occurred in the best-fit settings as well as the conditional probability that a value V1 of parameter P would appear in a best-fit region given that another value V2 of the same parameter P was present. Using these results, we were able to identify parameter values that virtually never appeared in any best-fit settings as well as those that only appeared if another value of the same parameter was also present. In either case, removing the parameter value would reduce the size of parameter space without significantly affecting the quality of fit (at least for these 20 subjects on these four task subsets).

6.3. Large and small parameter spaces for VR

Using this technique we computed two subsets of parameter space — one large and one small. In the large case, we only removed 14 parameter values, but this reduced the size of the parameter space by a factor of 200. These were values that the above analysis suggested could be safely removed and whose absence we intuitively thought would not be very important. The parameter values we removed are indicated by a minus sign in Figure 6-1. The analysis also suggested removing value "b" for the falsification parameter — parameter 22 (only 2% of our best-fitting settings contained that value), but as we will see later, the presence or absence of this value has important consequences. So we decided to keep it to allow us to further investigate falsification. In the small case, we tried to reduce the size of parameter space as much as possible while still capturing a lot of the individual differences. Again, we looked to the above analysis for suggestions about which values were most important as well as the most intuitively plausible. The parameter values we kept are indicated by a plus sign in the figure.

In the course of our analyses, we also decided to add an additional parameter to the small subspace to see if we could improve its fit while keeping it small. We noticed that VR was fitting tasks that did not involve "All" much better than those that did. Specifically, VR responded with NVC to these tasks more often than did the subjects. This phenomenon is reminiscent of Lee's *A-effect* — premise pairs involving "All" seem to facilitate the generation of conclusions (Lee, 1983). Inder (1987) argued that the effect could be explained in terms of a syntactic generation strategy for premise pairs involving "All". The basic idea is that given the premise *All y are x*, subjects may just respond with the other premise, after having replaced the y in it with x. For example, given the premise pair *All y are x, Some z are not y*, this strategy would replace the y in the second premise with x and produce the response *Some z are not x*. Notice that this strategy only applies if the middle term (y in this case), appears as the topic of an "All" premise. A more extreme version of this strategy would do the replacement even if the middle term only appeared as the predicate of an "All" premise. The parameter we added allowed for both strategies as well the absence of either:

```
23. All substitution strategy
  a. do not apply it
  b. apply if middle term appears as topic of "All"
  c. apply for any "All" premise
```

6.4. VR's fits to individual data

Figure 6-2 presents VR's average fits to the 103 subjects for both the large parameter subspace (without the all-substitution strategy) and the small parameter subspace (with it). We used complementary tasks for setting the parameters and evaluating the fit. Specifically, we chose four task subsets that contained 16 tasks, four that contained 32 tasks, and four that contained 48 tasks[14]. We also included the null set (0 tasks). Then for each subject, we computed all the best-fitting parameter settings for each subset of tasks. In the case of the null set, no parameter settings could fit better than any other, so all were included.

Given such a set of best-fitting settings, we then computed their average fit (i.e., the average proportion of tasks for which VR and the subject gave the same response) over all the *other* tasks (those that were not used to set the parameters) for which the subject gave a legal response. If VR predicted N responses for a task, one of which was the observed response, then the fit for that task would be 1/N (usually N was 1). The overall fit was the sum of these task fits divided by the number of tasks. For the large parameter space, it was not feasible to average over *all* best-fitting parameter settings in the 0 task case because this constituted searching the entire space. Instead, we took a random sample of 1000 parameter settings and averaged their fits on the 64 tasks. The figure presents the average of all these mean fits for each space and set of fitting tasks.

6.5. Comparisons with reference theories

Figure 6-3 presents several additional analyses that provide useful reference theories for comparison. On the far left is the expected percentage of accurate predictions for a completely random theory. Since there are nine legal conclusions, the probability of such a theory predicting any particular response is 11% (1 in 9). We also computed the number of identical responses between pairs of subjects and averaged over all such pairs. On average, 44% of responses were consistent across pairs of subjects and this is labeled

[14] The sets of 16 tasks are specified in footnote [12]. The sets of 32 tasks were formed by combining pairs of those 16 tasks — specifically, sets 1 & 2, 3 & 4, 1 & 4, and 2 & 3. Similarly, the sets of 48 tasks were formed by combining sets {2, 3, 4}, {1, 3, 4}, {1, 2, 4}, and {1, 2, 3}. Note that all 64 tasks are equally represented — each appears in one of the four sets of 16 tasks, in two of the four sets of 32 tasks, and in three of the four sets of 48 tasks. This is important because it means all 64 tasks are also equally weighted in the evaluation (described below).

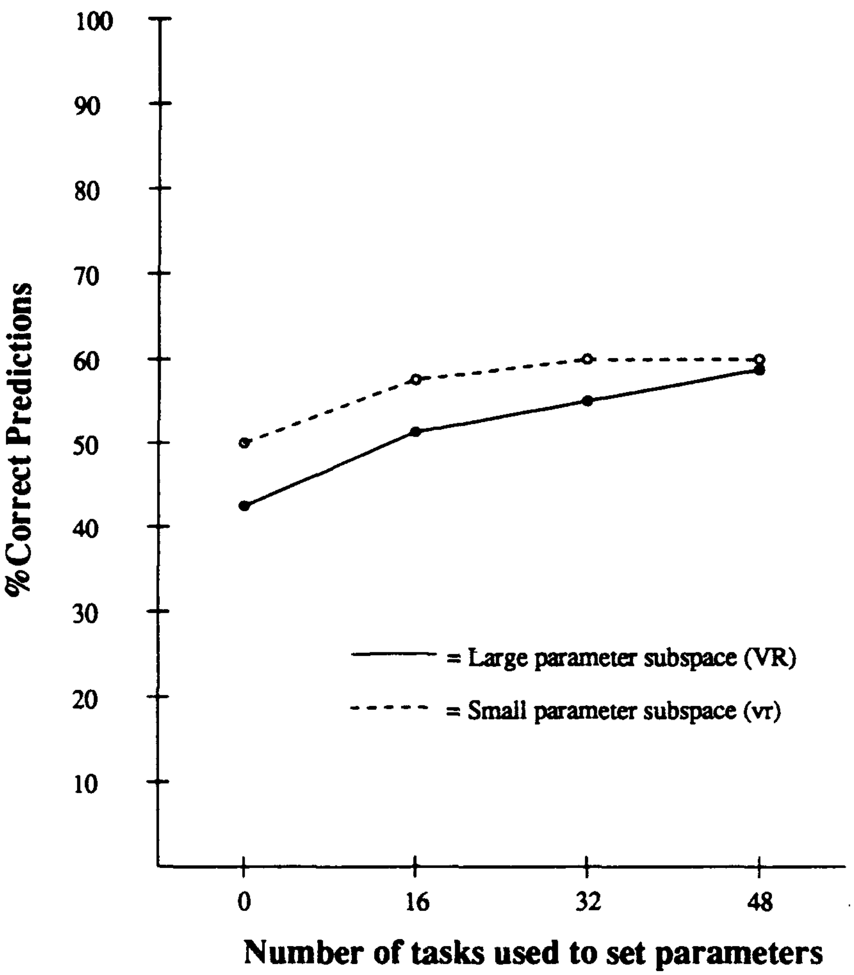


Figure 6-2: VR's fits to individual data.

other-subjects. The correct theory predicts correct performance on every task. For tasks in which there are multiple correct responses, this theory chooses randomly. Such a theory predicts 49% of the observed responses from the individuals in our data sets. Test-retest is the average proportion of responses given by a specific subject during one session that exactly match his or her responses to the same tasks a week later — on average 58% of responses are ide^tical. This is based on the one data set (Johnson-Laird & Steedman, 1978) that retested the same 20 subjects a weel later. VR's average fits to this same subset of subjects are given by Test-retest (VR) (59% — the large par ¬eter space) and Test-retest (vr) (62% — the small parameter space). The modal case (59%) is a post-hoc analysis in which the most common response over all the subjects is treated as the prediction. Finally, VR and vr correspond to VR's average fit to all 103 subjects for the large and small parameter space, respectively. All of VR's fits in this figure were computed using the four sets of 48 tasks to set the parameters and the other tasks to evaluate the fit.

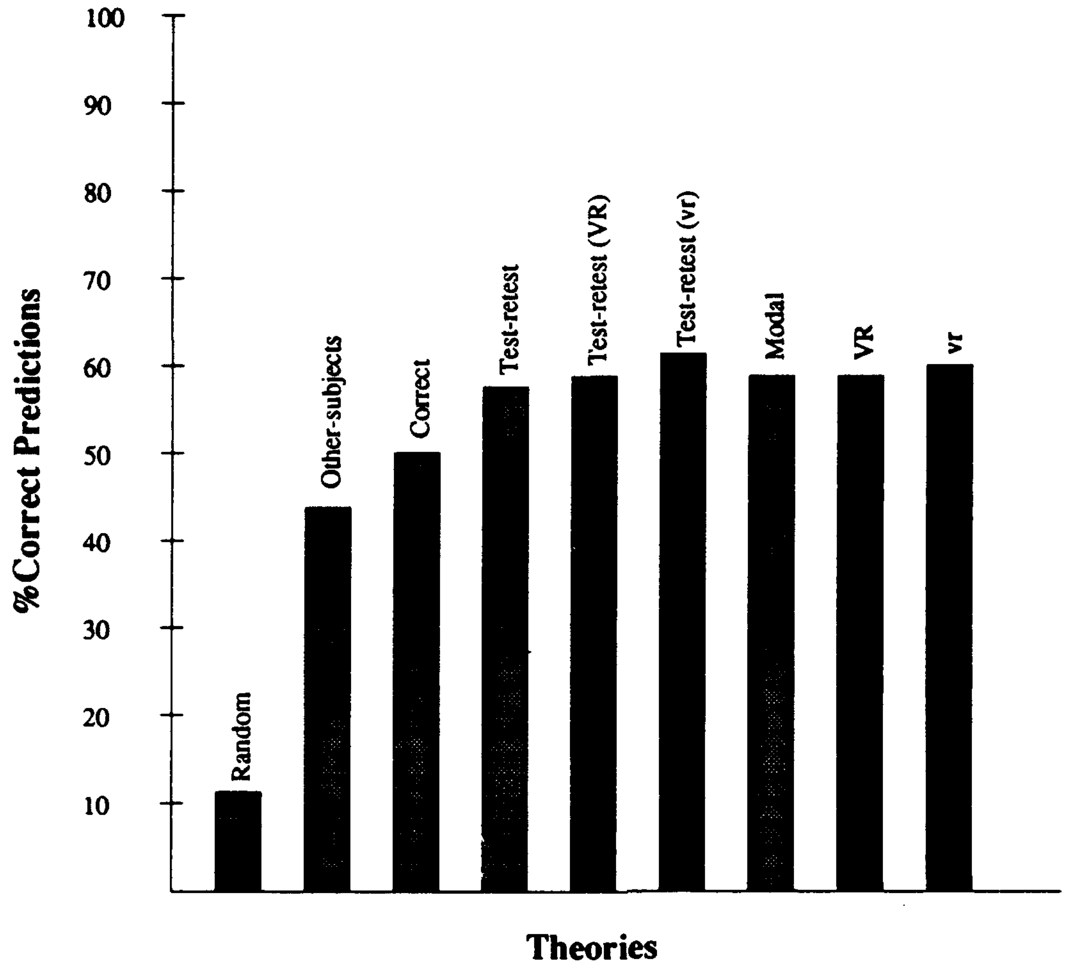


Figure 6-3: Comparison of VR's fits with various reference theories.

6.6. VR's fits to artificial data

In order to evaluate the range of behaviors admitted by VR, we computed its fits to some artificially generated data. These are presented in Figure 6-4. For random performance, we generated five data sets consisting of randomly generated responses to all 64 tasks. The fits presented are the average fit over these five for each parameter space. Correct performance consisted of correct responses to all the tasks. If a task had more than one correct response, predicting any of them was considered an accurate prediction. Modal performance corresponded to the behavior of the modal theory described above — for each task, we selected the most frequently observed response. Since the point of these analyses was to determine to what extent each type of behavior could be simulated by some parameter setting, we computed the true optimal fits over all 64 tasks, rather than over a subset of tasks as above.

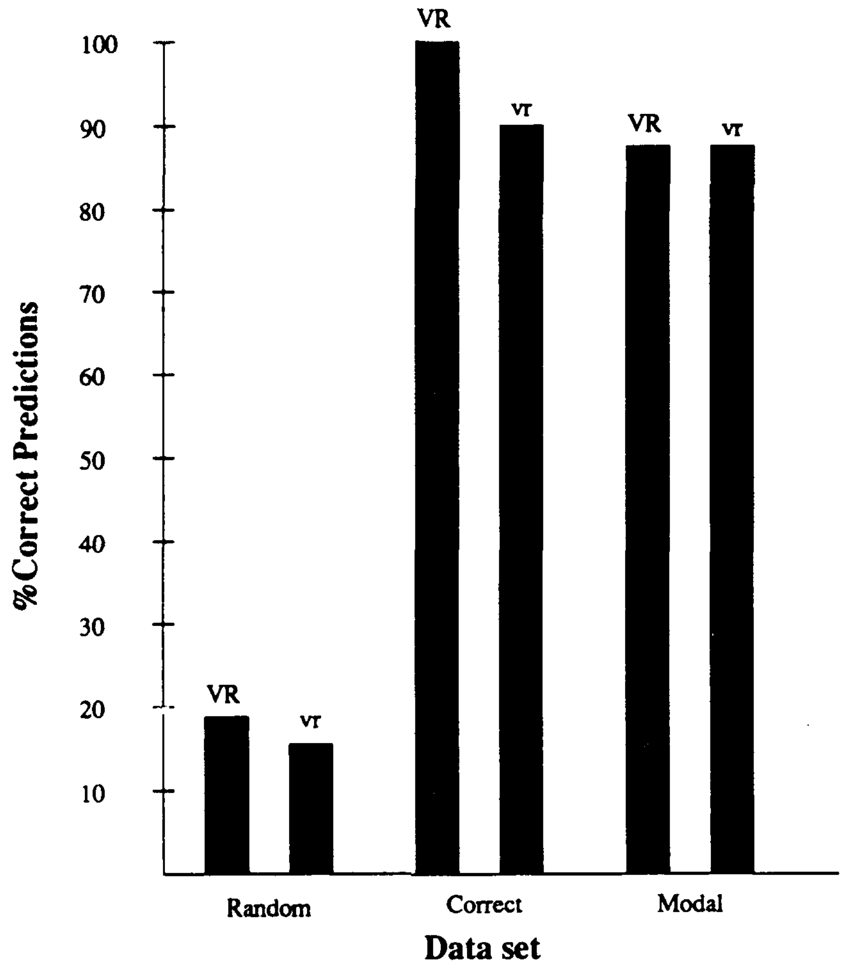


Figure 6-4: VR's fits to artificially generated data.

6.7. Analysis of falsification

Finally, Figure 6-5 presents an analysis of how well VR (the small parameter space) can fit individual data with and without falsification[15]. The top row in each pair presents the percentage of legal responses that were accurately predicted when falsification was not used (parameter 22 had value a), while in the bottom row falsification was used. All 64 tasks were used to set the parameters in both cases. The boxes indicate which fit was better. Heavy boxes indicate a disparity in the fit of more than 5%. The six pairs of rows

[15] It was not possible to obtain these fits for the large parameter space for computational reasons. ASPM (the tool we used to perform all our analyses) can usually reduce the fitting problem to a subset of parameter space. The nature of the additional falsification values, however, made this impossible and virtually the entire parameter space had to be searched. This was not feasible in the large space. See the discussion of *loosely-coupled tasks* in Polk, Newell, & VanLehn (1992) for a detailed treatment of the issue.

correspond to the six data sets from Figure 5-1. The first five have 20 subjects each, while the last (Inder) has only 3 subjects.

	Subject																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
a	48.9	$\boxed{54.7}$	$\boxed{46.7}$	$\boxed{67.3}$	$\boxed{52.5}$	$\boxed{68.3}$	58.1	$\boxed{60.9}$	$\boxed{66.7}$	$\boxed{58.1}$	$\boxed{60.0}$	$\boxed{50.0}$	$\boxed{50.9}$	$\boxed{54.7}$	31.7	$\boxed{75.0}$	$\boxed{58.7}$	$\boxed{82.5}$	54.7	$\boxed{65.0}$
bcd	$\boxed{60.0}$	50.0	41.6	59.2	45.9	61.9	$\boxed{61.3}$	57.8	60.3	56.5	56.7	48.4	41.5	50.0	$\boxed{36.7}$	71.9	55.6	81.0	$\boxed{56.3}$	61.7
a	61.3	$\boxed{64.5}$	$\boxed{69.8}$	$\boxed{80.6}$	$\boxed{85.9}$	74.2	$\boxed{76.6}$	79.7	$\boxed{48.4}$	57.8	$\boxed{66.7}$	$\boxed{64.5}$	38.7	$\boxed{63.3}$	$\boxed{80.0}$	59.4	65.6	$\boxed{84.4}$	$\boxed{68.8}$	$\boxed{49.2}$
bcd	$\boxed{62.9}$	56.5	66.7	77.4	84.4	$\boxed{75.8}$	75.0	$\boxed{82.8}$	45.3	57.8	65.1	56.5	$\boxed{41.9}$	61.7	73.3	$\boxed{62.5}$	65.6	81.2	62.5	46.0
a	66.7	$\boxed{62.5}$	$\boxed{82.8}$	$\boxed{79.4}$	$\boxed{81.3}$	66.7	$\boxed{59.4}$	87.5	$\boxed{43.8}$	$\boxed{70.3}$	61.9	$\boxed{67.2}$	$\boxed{30.2}$	$\boxed{42.1}$	$\boxed{68.3}$	$\boxed{68.3}$	$\boxed{85.9}$	$\boxed{85.9}$	$\boxed{67.2}$	$\boxed{48.4}$
bcd	66.7	60.9	79.7	76.2	79.7	$\boxed{68.3}$	56.3	87.5	42.2	67.2	61.9	62.5	28.6	38.6	66.7	66.7	82.8	82.8	62.5	46.8
a	55.7	$\boxed{63.5}$	$\boxed{63.5}$	$\boxed{50.0}$	81.3	$\boxed{77.4}$	$\boxed{64.5}$	73.4	75.0	60.9	68.8	$\boxed{72.6}$	62.5	$\boxed{65.1}$	$\boxed{79.7}$	54.2	$\boxed{61.9}$	$\boxed{47.5}$	$\boxed{76.6}$	62.9
bcd	55.7	58.7	55.6	46.8	81.3	75.8	59.7	$\boxed{75.0}$	$\boxed{76.6}$	60.9	68.8	71.0	62.5	58.7	68.8	54.2	58.7	44.3	75.0	$\boxed{66.1}$
a	66.7	81.3	$\boxed{67.2}$	$\boxed{56.3}$	77.8	77.8	69.4	78.1	$\boxed{75.0}$	$\boxed{75.0}$	78.1	$\boxed{67.2}$	72.6	79.7	$\boxed{76.6}$	$\boxed{58.7}$	64.5	$\boxed{61.3}$	77.4	60.9
bcd	$\boxed{69.8}$	81.3	65.6	54.7	$\boxed{79.4}$	$\boxed{79.4}$	69.4	78.1	71.9	73.4	$\boxed{79.7}$	65.6	72.6	79.7	73.4	57.1	64.5	54.8	77.4	$\boxed{64.1}$
a	$\boxed{73.0}$	68.9	57.8																	
bcd	71.4	$\boxed{70.5}$	57.8																	

Figure 6-5: VR's fits (small space) with and without falsification (bcd & a respectively).

Two additional values were added to the falsification parameter so that it now consisted of four values:

```
22. How to falsify
  a. don't
  b. do, if succeed then NVC
  c. do, if succeed then respond
     based on new annotated model
  d. do, recursively apply (c) above
```

The last two values correspond to falsification strategies similar to that proposed by Johnson-Laird (Johnson-Laird & Bara, 1984; Johnson-Laird & Byrne, 1991). With value c, VR attempts to construct an annotated model in which the premises are still true but not the conclusion. If it succeeds then it generates a new conclusion based on the resulting model (rather than responding NVC as value b does). With value d, VR recursively applies this strategy. That is, every time it generates a putative conclusion, it tries to construct an alternative model. If it succeeds and that new model supports yet another (different) conclusion, then VR attempts to falsify it. It continues in this way until it cannot falsify a conclusion or until the only conclusions it can generate have already been falsified (in which case it responds NVC).

6.8. Discussion

Perhaps the most surprising aspect of these results is that the smaller parameter space actually performs *better* than the larger space. In the four sets of fitting tasks in Figure 6-2, the smaller space produces fits that are between 2% (48 fitting tasks) and 8% (0 fitting tasks) better than the larger space. We believe there are two main factors that contributed to this unexpected result. First, the small space, but not the large space, included the all-substitution strategy. Consequently, the small space had an advantage in fitting subjects who applied this strategy despite the greater flexibility of the large space. Second, the added flexibility of the larger space could *reduce* the fit's quality because of overfitting — by including parameter settings that lead to rarely observed behaviors. For example, suppose we were using 16 tasks to set the parameters and the optimal fit on those tasks was 15 correct predictions. Then all and only the settings that produced that level of fit would be included in the analysis, regardless of how common or uncommon we might consider those settings a priori. Thus, plausible settings that produce a slightly lower fit (say 14 out of 16) would not be included, while implausible settings that just happened to do well on the set of fitting tasks would be. These implausible settings would presumably not do very well fitting the evaluation tasks and the average fit would go down. This problem is not nearly as severe for the smaller space because it only includes plausible settings. Furthermore, adding additional fitting tasks alleviates the problem because it becomes less and less likely that the implausible settings will fortuitously produce optimal fits on the fitting tasks. This explains why the disparity in fits between the large and small space is largest in the 0-task case (where none of the implausible settings are ruled out in the fit) and smallest in the 48-task case (where almost all of them are).

As expected the fits improve as additional tasks are used to set the parameters. When no tasks are used, then the fit just indicates how well the settings in the space do *on average*. As more fitting tasks are added, VR can fine-tune its behavior to model the individual subjects more closely. Also note that the first few fitting tasks lead to much larger improvements than do the last few — the difference between using 0 and 16 tasks is much larger than the difference between using 32 and 48. Apparently, the first few tasks rule out portions of parameter space that are way off the mark while the last few can only lead to small refinements in the predictions, not major improvements. Indeed, the smaller space appears to level off at around 60% correct predictions. Adding additional fitting tasks would probably not help much. So 60% correct predictions is a stable estimate of how well the smaller space can do in predicting individual behavior. It is important to note that in one sense these are *0-parameter fits* — none of the data used to evaluate the fits contributed to the setting of the parameters (because different tasks were used for each purpose). Consequently, overfitting is not a problem in these results.

The reference theories in Figure 6-3 provide a baseline against which to compare VR's fits. As one would expect, VR's predictions are significantly better than one would get by predicting randomly or assuming that responses were always correct. What is much more impressive is that VR's fits are at or beyond the test-retest reliability of the subjects themselves. Test-retest reliability gives an estimate of the stability and systematicity of the subjects and thus provides a pessimistic estimate of how well a fixed deterministic theory could possibly do without overfitting the data[16]. Consequently, one could hardly ask for more from VR without attempting to capture the instability of the subjects' performance over time (e.g., learning). Modeling instability in subject behavior is a natural next step for our approach but one that we have not yet undertaken.

The other-subjects and modal cases provide insight into how much of VR's fit could be due to capturing group trends vs. individual differences. The other-subjects result gives an estimate of how well a theory without any parameters could do if it only had access to one subject's data. Such a theory could be tailored to predict that one subject perfectly, but without any parameters to allow for individual differences only 44% of its predictions would be accurate on average. Clearly, VR is capturing significantly more of the behavior than this.

The best that any fixed theory (i.e., one without parameters) could possibly do would be to behave like the modal theory — to predict the most commonly observed response for every task. This theory's success (59% correct predictions) demonstrates what a large proportion of behavior could, in principle, be explained without reference to any individual differences. Of course, this theory provides an upper bound which is unlikely to be attained in practice. Nevertheless, it is worth pointing out that the modal theory's accuracy is comparable to the test-retest reliability of the subjects. That is, the most common response is as good a predictor of performance as is a specific subject's own behavior a week earlier. In any case, the fact that VR's predictions are as good or better than the modal theory shows that VR can predict behavior more accurately than any fixed theory could.

It is also interesting to compare the modal theory's performance with VP when the number of tasks used to set the parameters is zero (VR is being treated as a fixed theory, or set of fixed theories). In the case of the small parameter space (which was designed to include only the most common parameter settings), the quality of VR's predictions are within 10% of optimality.

Since the smaller space produces such good fits, the few parameters that it allows to vary account for a lot of the individual differences. We can identify these parameters in

[16] The estimate is pessimistic because the two tests occurred a week apart. If they had been administered together, the test-retest reliability estimate would probably be a little higher, although such a measure would also confound reliability with memory of the responses.

Figure 6-1 because they have more than one plus beside them (the others have a fixed value and cannot vary). These parameters can be grouped into two major classes: those involving the interpretation of particular premises (*Some* and *Some not* — parameters 1, 2, 4, and 6) and those involving the extraction of indirect knowledge about a proposition's predicate (parameters 10-13). In addition to these two, the all-substitution strategy (the new parameter) is another important source of individual differences — including more than one of its values (especially values "a" and "c") was important in obtaining good fits for different subjects. According to VR then, the three most important sources of individual differences in syllogistic reasoning (at least of those that we included) are: (1) the interpretation of the quantifier *Some*, (2) what indirect knowledge subjects can extract from propositions about their predicates, and (3) whether or not subjects apply the syntactic all-substitution strategy.

Figure 6-4 demonstrates some interesting points about the scope of VR — that is, what types of behavior it can and cannot fit well. First, note that with the large parameter space, VR is able to fit correct performance perfectly. In other words, there is at least one (actually many) settings in parameter space that lead VR to get all the tasks right — VR allows for perfect performance. Second, VR can fit modal data very well, but not perfectly. Perhaps if VR was a perfect theory it would be able to predict modal data completely. On the other hand, it is possible that *no* subject completely exhibits modal performance, and trying to predict it could lead to incorrect theories[17]. Third, VR is very bad at fitting random data. Less than 20% of its predictions are accurate on average and this is very close to what would be expected even if the parameters were not being fitted at all (11%). The slight improvement in fit can be attributed to overfitting — the data being used to set the parameters are also being used to evaluate the fit (recall that this was not a problem in Figure 6-2). The fact that VR cannot fit random data demonstrates its falsifiability. It is possible to observe data with which VR is inconsistent.

Notice that the larger parameter space fit the random and correct data better than the smaller space did. The reason is that all 64 tasks were used to set the parameters (rather than a subset of tasks as in Figure 6-2). Consequently, the larger space could exploit its greater flexibility to achieve better fits. Since the same tasks were used for both fitting and evaluating the parameter settings in these analyses, it did not run the risk of finding settings the performed well on the fitting tasks, but poorly on the evaluation tasks. Nevertheless, the larger space did not fit modal data any better than did the smaller space. Apparently, the addition of the all-substitution strategy made up for the smaller space's reduced flexibility in this case.

[17] For a compelling discussion of some of the dangers of trying to fit aggregate rather than individual data see Siegler (1987).

Finally, the results in Figure 6-5 illustrate a number of points about falsification. First, there is not a very large difference in the fits with and without falsification. Only 15 subjects' fits (out of 103) were affected by more than 5% based on this parameter. Furthermore, in only one case out of 103 was the fit *with* falsification more than 5% better than the fit without. This result is particularly striking given that the parameter space was three times *larger* with falsification than without it (values b, c, and d were all considered consistent with falsification), allowing for a wider range of behavior. Clearly then, falsification is an unnecessary assumption that only weakens VR as a theory. The main reason we included it was that it is central to Johnson-Laird's theories of reasoning so we wanted to see whether it would improve VR's fits. The results in Figure 6-5 demonstrate that it did not.

Chapter 7
General Discussion

VR provides an accurate computational theory of behavior on categorical syllogisms. Not only can it be run to produce the major regularities that have been discovered, but it can also be used to generate other testable predictions and to predict the behavior of individual subjects with surprising accuracy. Given such a theory, we are now in a position to discuss some of the central issues surrounding human reasoning. For example, are human beings rational? If so, why do they make so many errors in straightforward reasoning tasks? And what is the best way to characterize the processes people apply when reasoning? By analyzing VR's structure, we hope to gain some insight into these questions.

7.1. Verbal reasoning

Despite their differences, all previous reasoning theories have shared the same basic structure at the most general level — they *encode* the problem into an internal representation, *reason* using operations on that representation, and *decode* the result to produce a response. Different hypotheses have been put forth concerning the nature of the reasoning operations including searching for alternative mental models (Johnson-Laird & Byrne, 1991; Johnson-Laird & Bara, 1984), applying logical inference rules (Braine, 1978; Rips, 1983), and invoking context-specific inferences based on schemas (Cheng & Holyoak, 1985), but they are all based on the same assumption — the cognitive mechanisms that are most important in reasoning are devoted primarily (or exclusively) to that task.

The structure of VR suggests a very different answer. After initial encoding, VR repeatedly reencodes the premises until it can generate a legal conclusion (see Figure 3-1). But all these processes (initial encoding, reencoding, and conclusion generation) are fundamentally linguistic and are not devoted primarily to reasoning. VR's next process, giving up on reencoding, is also clearly needed to avoid excessive delays in other tasks (e.g., reading), although one could argue that the exhaustiveness of reencoding in VR would be unique to reasoning. In any case, giving up is certainly much less central to VR's behavior than are reencoding and conclusioı. generation.

Falsification is the one process in VR that is clearly designed specifically for reasoning. Furthermore, when used, its role could be every bit as important as that of the other processes — especially if a sequence of putative conclusions are successively falsified, leading to a completely different response. As mentioned above though, falsification proved to be an extraneous assumption in achieving the fits we did with VR. Only one subject's fit (out of 103) improved significantly when VR falsified as opposed to when it did not. So we have no reason to believe that untrained subjects like those we studied attempt to falsify their conclusions. In short, VR suggests that the most important processes in syllogistic reasoning are linguistic — especially reencoding — and that reasoning-specific processes play a much smaller role by comparison.

A natural question is whether syllogistic reasoning is really just an elaborate form of language comprehension according to VR. The answer is no. Although it's true that encoding and reencoding the premises are at the heart of its behavior, VR's control structure reflects the demands of syllogistic reasoning rather than language comprehension. For example, if generation fails to produce a legal conclusion, then reencoding is repeatedly evoked until every possible reference property has been tried. Exhaustive reencoding is a response to a demand of the syllogism task (producing a conclusion that relates the end terms) and would be inappropriate in typical comprehension. Similarly, conclusion generation only succeeds if the candidate conclusion is a legal syllogism response. Again, this constraint comes from the syllogism task and would not apply in other domains.

In short, we believe that most of the "reasoning" on syllogisms is done using standard linguistic processes, but that these processes are used in a way that reflects the demands of the syllogism task. We refer to such behavior as *verbal reasoning* (hence the system's name — VR for verbal reasoner). More precisely,

> **Verbal reasoning is the deployment of linguistic processes according to and in order to satisfy the demands of a reasoning task.**

A major claim of this dissertation is that the behavior of untrained subjects on categorical syllogisms can best be characterized as verbal reasoning.

7.2. Additional evidence in favor of verbal reasoning

Of course, the main evidence in favor of verbal reasoning is VR's success in modeling human data. But there are also ecological considerations that recommend it. For most human beings, communication is a far more common and important task than deductive reasoning. We spend far more time talking, listening, reading and writing than we do trying to generate logically valid conclusions from premises or testing the validity of arguments. Given such an environment, one would expect the cognitive processes that have been developed for communication to be far more sophisticated than those for deductive reasoning (if any have been developed specifically for the domain at all).

Under such circumstances, verbal reasoning would be a natural and very rational response when faced with a categorical syllogism. In addition to any available reasoning processes (to ensure validity), the task clearly calls for linguistic processes to encode the premises (which are, after all, linguistic) and to produce a verbal response. And while the linguistic processes are extremely sophisticated, we are assuming that the reasoning processes are either underdeveloped or missing entirely. Attempting to adapt well-developed linguistic processes to the demands of the syllogism task is probably the most reasonable way for an untrained person to make progress on the task, if not the *only* way.

But even if verbal reasoning is a rational strategy for solving syllogisms, that does not imply that it will lead to perfect performance. The problem is that adapting linguistic processes to the syllogism task cannot be done instantaneously. Like most highly skilled processes, encoding, reencoding, and generation processes are undoubtedly largely automatic — people do not have deliberate control over their internal workings. Subjects can organize the flow of control through these processes to meet the demands of the task (as described above), but they have to take what they get whenever one is evoked. And while these outputs serve the needs of communication almost perfectly, they are less well suited to the demands of deductive reasoning. For example, VR assumes that (re)encoding produces an annotated model that will be *consistent with the premises*, but that may be unwarranted from the standpoint of logic. In standard communication, however, people often expect the listener to make plausible (if invalid) assumptions without being told explicitly. For example, if someone asks you if you know what time it is, it goes without saying that they want you to tell them (rather than just answer yes or no). Similarly, VR's assumption that direct knowledge (about topics) is more available than indirect knowledge is better adapted to the needs of communication than deductive reasoning. If a speaker wants to convey a fact about dogs then he will tend to use "dogs" as the topic of his statement — "no dogs are cats" rather than "no cats are dogs" although logically the two are equivalent. In short, the responses people make to syllogisms are decidedly rational — they are based on an attempt to make use of sophisticated linguistic processes to satisfy the demands of the task. The only reason the behavior appears irrational at times is that those linguistic processes are adapted to the needs of communication rather than syllogistic reasoning[18].

Verbal reasoning is a parsimonious account of syllogistic reasoning — it accounts for aggregate and individual reasoning data mainly in terms of standard linguistic processes,

[18] This argument was inspired by similar accounts made by Anderson (1990). He argues that we can understand a lot about cognition by considering what abilities are most important given the nature of the environment — what he calls a "rational analysis".

rather than reasoning-specific mechanisms. It does not follow, however, that verbal reasoning rules out such mechanisms. Verbal reasoning is essentially a theory of subjects who have not had any training in logic or reasoning. None of the 103 subjects we analyzed had had experience with these types of problems before, and hence they presumably lacked any task-specific skills. But people can certainly be trained in alternative strategies to improve their performance. Venn diagrams and Euler circles are examples of approaches that are regularly taught to improve reasoning performance. Obviously, VR is not intended to model subjects using such strategies. Given enough time, subjects could even begin to modify their linguistic processes in task-specific ways[19]. For example, they could learn that for these tasks they should interpret "Some x" as referring to a set of x's that are distinct from any other x's in the problem. Such an interpretation would help subjects avoid the fallacy of the undistributed middle term (described on page 41) and improve performance. But developing new strategies and modifying automated processes takes practice. Without such practice, we assume subjects use verbal reasoning.

7.3. The methodology behind verbal reasoning

The idea of deploying well-developed processes according to the demands of a given task may also provide assistance in developing plausible theories of other high-level cognitive behavior. Let us refer to a well-practiced or automated cognitive process as a *skill*. Given a task, one can look to the environment to suggest skills that are both adaptive in the environment and relevant to the task. The demands of the task can then provide constraint on organizing those skills together into a theory that has the functional capabilities to perform the task.

This methodology provides the hope of leading to theories that have both independent plausibility and predictive power. Choosing skills that are adaptivr in the environment provides constraint on the major constructs in the theory and gives it plausibility independent of its empirical success. Furthermore, an analysis of how well or poorly those skills are adapted to the task of interest should help in predicting errors as well as correct performance.

Of course, there's no simple recipe for developing good theories. The above methodology will undoubtedly help in some cases but not in others. Nevertheless, it does reflect the basic view from which our more general ideas about verbal reasoning arose and so it is worth making explicit. We also hasten to point out that although these ideas arose out of our own work on verbal reasoning, they are very similar to and were undoubtedly influenced by Anderson's (1990) earlier work on rational analysis.

[19] Lehman, Newell, Polk, & Lewis (1992) refer to this process as the "taskification" of language and discuss it in much greater depth.

7.4. The relationship between VR and Soar

Finally, given that earlier versions of the theory were implemented in Soar, a cognitive architecture which should be familiar, it is worth describing the relationship between Soar and VR. As mentioned in Section 2.3, VR is implemented in Lisp in order to increase its speed, but it could certainly have been implemented in Soar (previous versions of the theory were). In fact, Soar would have allowed us to use the more flexible control structure afforded by production systems and which is so characteristic of human behavior. Given that Soar has been proposed as a unified theory of cognition (Newell, 1990), two natural questions arise: does VR's success in modeling syllogistic reasoning provide evidence for Soar as a unified theory of cognition, and conversely, does Soar's success in modeling cognitive behavior on other tasks provide additional evidence for VR?

Since VR could have been implemented in Soar, the assumptions built into Soar and VR are clearly consistent with each other. The success of each thus provides *some* evidence for the other, although it is weak at best. But the fact that VR is implemented in Lisp demonstrates both that Soar's basic assumptions (e.g., problem spaces, universal subgoaling, learning by chunking) are not essential in deriving VR's predictions and that VR's assumptions are not essential to Soar's success in other domains.

The problem is that Soar and VR are theories of different levels of behavior. Soar is a theory of the cognitive architecture — the set of fixed mechanisms that underly all cognitive behavior — while VR is a theory of behavior on a specific high-level cognitive task. It is understandable then why the two do not provide much mutual constraint — the assumptions that provide predictive power for a high-level task such as syllogistic reasoning (consistent encodings, availability of direct knowledge, ...) are themselves high-level rather than architectural. Put simply, high-level behavior can often be explained without reference to the architecture.

Nevertheless, there is at least one way to make use of a cognitive architecture in providing constraints on high-level theories. The idea is to develop a range of theories within the same architecture that share some common skills. For example, if VR used the same linguistic processes used in accurate theories of two other tasks (preferably using identical computer code), then all three would provide mutual support for each other. Furthermore, if some of the shared skills depended on architectural assumptions (e.g., about learning, attention, ...), then the success (or failure) of the group of theories could also provide information about the architecture. Note that the architecture does not have to be Soar. The point is that by implementing multiple theories within *some* unified theory of cognition, they can mutually constrain each other and the architecture itself. Currently, however, VR and Soar do not provide much support for each other.

7.5. Description and evaluation of other theories

We now briefly describe seven other theories that have been proposed to account for behavior on categorical syllogisms and compare them with VR. Except for two classic theories whose absence would completely misrepresent the field (Woodworth & Sells (1935) and Chapman & Chapman (1959)), we will only discuss theories that (1) were intended to explain most or all of the basic processes in syllogistic reasoning, (2) have actually been used to explain some of the regularities listed above, and (3) appeared in books or refereed journals. Numerous articles have presented ideas about a small subset of the processes involved in syllogistic reasoning but were not meant to be complete theories (e.g., Begg & Harris (1982) proposed a theory of premise interpretation; Dickstein (1978) explained the figural effect). Other authors have described theories that might apply to categorical syllogisms, but have not actually used their theories to explain the specific regularities in this task (e.g., Henle, 1962; Clark, 1969; Braine, 1978; Rips, 1983). All these papers are clearly relevant in understanding behavior on categorical syllogisms, but since they were not meant to be complete theories of the task, it would be both unfair and inappropriate to include them in the comparison. After briefly describing each theory, we will discuss potential problems and compare them with VR.

7.5.1. The atmosphere hypothesis

Woodworth & Sells (1935) were two of the first psychologists to attempt to understand human behavior on categorical syllogisms. They proposed a theory, known as the atmosphere hypothesis, to account for a subset of errors that people make. Begg & Denny (1969) provided a concise characterization of this hypothesis:

1. If either premise is particular (*Some* or *Some not*), then the conclusion will tend to be particular. Otherwise, the conclusion will tend to be universal (*All* or *No*).
2. If either premise is negative, then the conclusion will tend to be negative. Otherwise, the conclusion will tend to be positive.

The basic idea behind this hypothesis is that the premises produce a "set" in the subject reading them, and that this set biases the conclusions produced. As might be expected for the first theory of syllogistic reasoning, the atmosphere hypothesis is not as complete or accurate as more recent theories, including VR. For one thing, it was only intended to be a theory of errors. As such, it does not specify how people get syllogisms *right* (at least for syllogisms whose correct conclusion is not atmospheric), only how they get them wrong. Furthermore, it cannot account for NVC responses and these often constitute a significant portion of those observed. It also fails to predict the figural, belief bias, and elaboration effects. VR, on the other hand, provides a complete process model for both correct and incorrect responses (whether NVC or not) and computationally models all of the major regularities.

7.5.2. Theories based on conversion

Chapman & Chapman (1959) argued that many errors could be explained by assuming that subjects had illicitly converted one (or both) of the premises — assuming that *All x are y* leads to *All y are x* and that *Some x are not y* leads to *Some y are not x*. Revlis (1975b) presented a better specified theory based on the same assumption. Both these theories account for a number of errors but fail to provide a process model of correct performance — they assume the existence of what Revlis calls "an unspecified deduction operation". Empirically, neither theory predicts the belief bias effect. A more serious problem is that by assuming that the premises are converted, these theories predict that a syllogism's figure should not affect behavior (the only difference between figures is the order of terms and converting the premises neutralizes that difference). Consequently, without some modification, these theories are inconsistent with the figural effect. Again, VR provides a complete process model for both correct and incorrect responses and predicts all of the major regularities including the figural effect.

7.5.3. Theories based on Euler circles

Erickson (1974), Guyote & Sternberg (1981), and Fisher (1981) all proposed that subjects solve syllogisms by manipulating representations analogous to Euler circles. These theories assume that subjects encode each premise into Euler circles, combine these representations together (sometimes selecting a subset of the initial encodings), and then read out the conclusion from the result. These accounts are relatively complete — they provide detailed explanations of all correct and incorrect answers as well as NVC responses. But since all terms have equal status in an Euler circle, these theories do not always predict the figural effect (which depends on the end terms being distinguished in some way)[20]. Furthermore, none of these theories has modeled the belief bias effect. In contrast, VR has been used to model both phenomena.

7.5.4. Mental model theory

Johnson-Laird's mental models theory (Johnson-Laird & Bara, 1984; Johnson-Laird & Byrne, 1991) shares the most similarities with VR and so we will discuss it in some detail. Johnson-Laird & Bara (1984) outlined their theory of categorical syllogisms in terms of three basic steps (p. 5):

[20] By assuming that each conclusion is associated with a specific Euler circle configuration, these theories can occasionally predict a figural effect (if the figural conclusion matches the configuration, while non-figural conclusions do not). But in some cases, syllogisms in opposing figures are assumed to produce the *same* configuration (e.g., *No A are B, No B are C* and *No B are A, No C are B*). Clearly, these theories cannot predict figural effects for both tasks.

1. Construct a mental model of the premises, i.e. of the state of affairs they describe.
2. Formulate, if possible, an informative conclusion that is true in all models of the premises that have so far been constructed. An informative conclusion is one that, where possible, interrelates terms not explicitly related in the premises. If no such conclusion can be formulated, then there is no interesting conclusion from syllogistic premises.
3. If the previous step yields a conclusion, try to construct an alternative model of the premises that renders it false. If there is such a model, abandon the conclusion and return to step 2. If there is no such model, then the conclusion is valid.

The authors implemented the central tenets of this theory in computer programs and predicted that two main factors should affect the difficulty of a syllogism: the figure of the premises and the number of mental models that have to be constructed. In their theory, the figure of the premises affects the difficulty of constructing an initial model and of formulating a conclusion whose subject appeared before its object in the premises. The number of models is important since each model makes demands on working memory. If the models required of a task exceed the capacity of working memory, then responses will be based on a subset of the models and responses will tend to be erroneous. Consequently, tasks that only require one model are predicted to be easier than those that require more. Johnson-Laird & Bara ran experiments that both confirmed these predictions and showed that their computer program could account for the most frequent responses on almost all 64 syllogism tasks.

The Johnson-Laird theory is probably the most successful of any of the previous theories. It explains correct as well as incorrect responses (whether NVC or not), and has been use^d to explicitly predict three of the major phenomena (difficulty, figural, and belief bias effects). Furthermore, the specific predictions of the theory on the 64 tasks show evidence for three of the other four (validity, atmosphere, and conversion effects). Finally, although the elaboration effect has not been explicitly modeled, the theory also clearly predicts it, since, the theory assumes that the mental model must be consistent with the premises. Consequently, like VR, elaborating the premises to be unambiguous will constrain the mental model to be closer to valid, leading to better performance.

7.6. The relationship between VR and mental model theory

It should be clear that VR was significantly influenced by Johnson-Laird's theory — the annotated models used in VR were based directly on his mental models (with some slight modifications). It may be tempting to suppose that VR is really just a variant of Johnson-Laird's approach — that it doesn't represent an alternative theory at all. But there is a fundamental difference in how the two theories view the process of syllogistic reasoning. In Johnson-Laird's theory, the heart of the process is the search for alternative models —

a process that is devoted primarily, if not exclusively, to reasoning. In contrast, the central processes in VR are fundamentally linguistic, and are not reasoning-specific. In a very real sense then, Johnson-Laird's theory is more similar to other reasoning theories (even rule-based theories) that fit into the general *encode-reason-decode* structure than it is to VR.

7.7. Evidence favoring VR over mental model theory

Given that the theories do represent very different alternatives, how shall we choose between them? First, consider the plausibility of the two accounts. According to Johnson-Laird's theory "reasoners make assumptions, not by employing mental rules of inference, but rather by searching for alternative models of the premises that render putative conclusions false" (Johnson-Laird & Bara, 1984, p. 36). But such a strategy seems implausibly sophisticated for untrained subjects. Not only must they appreciate the need to consider as many alternatives as possible, but they must also possess the skills to modify an existing model in such a way as to falsify their putative conclusion *without* falsifying the premises. Furthermore, in the absence of working memory limitations, these skills must be *guaranteed* to produce an alternative model if one exists, since otherwise people would not be able to solve syllogisms correctly even in principal. Perhaps subjects could be taught to use such a process, but to assume that subjects possess it *without any training* seems implausible. VR's assumption that people reason using linguistic processes seems much more likely. No one would doubt the existence of sophisticated processes for language comprehension and generation and VR demonstrates the feasibility of employing them to do syllogistic reasoning.

The empirical analyses in this dissertation support this assessment. First, the falsification analysis showed that falsification was an extraneous assumption in achieving VR's fits. Only one subject's fit (out of 103) improved by more than 5% when VR falsified compared with when it did not. It is important to stress that the data did not *rule out* the search for alternative models — just that they were consistent with the more parsimonious hypothesis that subjects do not perform such a search. In terms of aggregate data, both theories can predict all the regularities (though Johnson-Laird has not *implemented* versions that produce belief bias and elaboration effects, he presumably could). But VR makes a number of other accurate predictions (Figure 5-8) and it remains to be seen whether these can be derived from Johnson-Laird's theory. Most importantly, VR accurately models the detailed behavior of individual subjects. Johnson-Laird's theory has never been used to predict data at such a fine level of detail and whether it can do so is an open question.

7.8. Conclusion

The traditional view of syllogistic reasoning assumes that subjects possess reasoning-specific processes for making inferences. Verbal reasoning offers a parsimonious alternative — untrained subjects solve syllogisms by deploying linguistic processes according to and in order to satisfy the demands of the syllogism task.

VR demonstrates that a detailed theory based on this idea is not only feasible, it actually predicts human behavior more accurately and in far more detail than previous theories. Furthermore, a simple ecological analysis suggests that without training verbal reasoning is probably the most rational approach to the task, if not the only possible one.

Verbal reasoning also offers the hope of explaining behavior on other reasoning tasks that involve primarily linguistic material. It is to such tasks that I now turn my attention.

Chapter 8
Other Tasks

8.1. Introduction

I argued in the last chapter that the mental model theory of syllogistic reasoning (Johnson-Laird & Bara, 1984; Johnson-Laird & Byrne, 1991) is the most successful of all previous accounts in capturing the empirical results. But there is an even more important reason for preferring it to earlier theories — unlike the others it has been used to account for behavior on many other reasoning tasks. In fact, in their recent book *Deduction*, Johnson-Laird & Byrne (1991) have developed mental model theories for *all* the standard tasks used in studying deductive reasoning. As one would expect, each individual microtheory makes a few task-specific assumptions, but they are all based on the same general framework. In short, the authors presented the first unified theory of deduction to date and this accomplishment provides significant support for their theory.

As discussed in Section 7.5.4, mental model theory assumes three major stages in reasoning: *comprehension* (encoding the problem statement into an initial model), *description* (formulating a conclusion based on the model), and *validation* (searching for an alternative model that falsifies the putative conclusion). I have argued that the fundamental difference between this theory and verbal reasoning is the search for alternative models (validation). Indeed, without validation, model theory reduces to the linguistic processes of comprehension and description — an example of verbal reasoning. But according to model theory, the search for alternative models is at the very heart of deductive reasoning. In contrast, verbal reasoning assumes that encoding and reencoding are the central processes. VR's success without validation demonstrates that by making certain assumptions about two of the three stages assumed by mental model theory (comprehension and description), the third (validation) becomes extraneous (at least for syllogistic reasoning). A natural question is how important validation is in explaining behavior on other tasks.

I will investigate this question by examining the microtheories proposed by Johnson-Laird & Byrne (1991) for each of the other deductive reasoning tasks. I will show that validation is unnecessary in explaining the empirical data addressed by the authors and that, as a result, most of their explanations correspond to verbal reasoning accounts.

Consequently, such an analysis will provide additional support for the verbal reasoning view of reasoning. It will demonstrate that verbal reasoning generalizes beyond categorical syllogisms to other deductive reasoning tasks. It will also identify behaviors that require assumptions beyond verbal reasoning and will thus provide some insight into the scope of the theory. Following Johnson-Laird & Byrne's organization, I will first consider propositional reasoning. I will then focus on conditional reasoning — a specific type of propositional reasoning that has been extensively studied. Next, I will discuss relational and quantificational reasoning and finally turn to meta-deduction.

8.2. Propositional reasoning

Chapter 3 of *Deduction* is devoted to propositional reasoning. Propositional reasoning tasks involve statements that can be assigned a truth value (*propositions*) and connectives (such as "and", "if", "or", and "not") for composing them into new propositions. Psychologists have mainly focused on two types of propositional reasoning: *conditional reasoning* (involving "if") and *disjunctive reasoning* (involving "or"). Figure 8-1 gives typical examples of both types of tasks.

Conditional Reasoning If p then q, p only if q		**Disjunctive Reasoning** p or q
Modus ponens (MP):	p, therefore q	-p, therefore q
Modus tollens (MT):	-q, therefore -p	-q, therefore p
Denied antecedent (DA):	-p, therefore -q	p, therefore -q
Affirmed consequent (AC):	q, therefore p	q, therefore -p

Figure 8-1: Propositional reasoning tasks.

Like categorical syllogisms, both tasks usually consist of two premises. One premise contains a propositional connective (*if p then q, p only if q, p or q*) while the other is just an atomic proposition or its negation (*p, q, -p, -q*). In addition to the connective "if", a conditional premise (*if p then q*) consists of an *antecedent* (*p* in *if p then q*) and a *consequent* (*q* in *if p then q*). There are four standard inferences that subjects tend to draw on conditional reasoning tasks (aside from "nothing follows"): *modus ponens (MP), modus tollens (MT), denied antecedent (DA),* and *affirmed consequent (AC)* (Figure 8-1, left). The first two are deductively valid while the last two are not. Similarly, there are four standard inferences on disjunctive tasks (Figure 8-1, right) and only two are deductively valid (the first two in the figure). Disjunctions are consistent with both an *inclusive* and *exclusive* interpretation. With an inclusive interpretation, the disjunction is considered true if either *or both* of the constituent propositions is true (*p or q, or both*). An exclusive disjunction is true only if one *but not both* of the propositions is true (*p or* *q, but not both*). With an exclusive interpretation, all four of the inferences on the right of Figu:e 8-1 are deductively valid. Many studies have added the words "or both" or "but not both" to make the intended interpretation unambiguous.

Now consider Johnson-Laird & Byrne's (1991) explanations of behavior on these tasks. Again, the goal is to assess the role of validation in these explanations. The first findings that Johnson-Laird & Byrne address involve the interpretation of disjunctions and conditionals. Neither type of proposition is consistently interpreted by subjects. Some experiments find a bias toward interpreting disjunctions as inclusive (Evans & Newstead, 1980) while others find a bias toward exclusive interpretations (Manktelow, 1980). Similarly, conditionals are sometimes interpreted as bi-conditionals (*if and only if p then q*) while at other times (even by the same subjects) they are nct. Johnson-Laird & Byrne explain these findings by assuming that subjects construct mental models that do not explicitly represent all the available information. For a disjunction such as *p or q*, the authors assume that subjects construct two models — one in which p is true and one in which *q* is true. Such a representation leaves open whether or not both *p* and *q* could be true at once (in the same model). Hence, it is consistent with both an inclusive and exclusive interpretation. For a conditional such as *if p then q*, Johnson-Laird & Byrne assume that subjects construct a singl- explicit model in which *p* is true and therefore *q* is true as well. This representation does not make explicit whether there could be other models in which *q* is true, but not *p*, and is consequently consistent with both a standard conditional interpretation and with a bi-conditional interpretation. The important point here is that neither of these explanations makes reference to validation. As long as comprehension delivers mental models like those assumed by Johnson-Laird & Byrne, then conditionals and disjunctions will be interpreted in an indeterminate way.

It is not particularly surprising that the previous explanations did not make reference to validation. After all, they were explanations of how subjects *interpreted* propositions — not how they reasoned with them. The other findings about propositional reasoning that Johnson-Laird & Byrne address, however, involve the diffic lty of different propositional inferences and so one might expect validation to play an important role. But even in these cases, as we will see, validation is not necessary to the authors' accounts.

Consider the finding that modus ponens is easier than modus tollens in the standard conditional reasoning task (Figure 8-1, lett). According to Johnson-Laird & Byrne, comprehending *if p then q* delivers two models — one in which both *p* and *q* are true, and one that has no explicit content. Since *p* is true in the first model, it accommodates the categorical premise for m..dus ponens (*p*) and the second model (with no explicit content) is eliminated. This model supports the inference that *q* is true (the standard modus ponens conclusion) and so it should be relatively easy. In contrast, the categorical premise for modus tollens (*-q*) leads to the elimination of the first model, leaving only the model with no explicit content (which is updated to reflect that *-q* is true). Since this model does not support the modus tollens conclusion (-*p*), making such an inference should be relatively hard. Once again though, note that this explanation makes no reference to validation. The critical assumptions involve the results of comprehension (whether the resulting models support the conclusion of the inference or not), not the search for alternative models.

The accounts of the other propositional reasoning results similarly depend on assumptions about comprehension rather than validation. The difference in difficulty of modus ponens and modus tollens is predicted to disappear when propositions of the form *p only if q* are used, because comprehension is assumed to deliver two explicit models that together support both inferences. And since dealing with two explicit models is assumed to be harder than dealing with one, both inferences are predicted to be harder than modus ponens is with the standard conditional, *if p then q*. Similarly, modus tollens is predicted to be easier with a bi-conditional (*if and only if p then q*) than with a standard conditional because the former is assumed to require only two explicit models while the latter is assumed to require three. No such difference is predicted for modus ponens since in both cases it is assumed to require only one explicit model. Conditionals are predicted to be easier than exclusive disjunctions because they call for the initial construction of only one explicit model, while exclusive disjunctions call for the construction of two. Finally, double disjunction tasks such as the following are predicted to be extremely difficult:

```
Linda is in Cannes or Mary is in Tripoli, or both.
Mary is in Havana or Cathy is in Sofia, or both.
What, if anything, follows?
```

The reason is that they are assumed to require the initial construction of a very large number of models (five in the above task). But if exclusive disjunctions are used instead of inclusive disjunctions (which require an additional model to represent the "or both" situation), the task is predicted to become easier.

The data bear out all these predictions. But notice that all these explanations are based on the number of explicit models that are assumed to be delivered by comprehension or on whether or not an initial model supports an inference. The important point here is that *none* of the accounts make reference to the construction of alternative models to falsify a putative conclusion — validation is extraneous in explaining these findings.

8.3. Conditional reasoning

Conditional reasoning has been extensively studied and many variants of the basic task presented in the last section have been formulated, leading to the discovery of many additional empirical results. In Chapter 4 of *Deduction*, Johnson-Laird & Byrne turn their attention to these other findings and argue that mental model theory can also account for them. Once again, my intent will be to show that the explanations that the authors provide for these results do not depend on validation.

The meaning of conditionals

Johnson-Laird & Byrne begin Chapter 4 by developing a model theory for the meaning of conditionals. They assume four types of situations: actual states of affairs (which are what really happened), real possibilities (which could happen given the actual state of the world), real impossibilities (which could never happen given the actual state of the world), and counterfactual situations (which once were real possibilities, but no longer are because they did not occur). They summarize their theory as follows (pp. 72-73):

1. An indicative conditional is interpreted by constructing an explicit model of its antecedent, which is exhaustively represented, and to which is added a model of the consequent. An alternative implicit model allows for cases in which the antecedent does not hold.
2. A counterfactual conditional is interpreted in the same way except that the models of its antecedent and consequent are of counterfactual situations, and there is an explicit model of the actual situation.
3. Conditionals may elicit richer models in which more states are rendered explicit. This fleshing out of models occurs in several circumstances, e.g. when a referential relation, or one based on general knowledge, holds between antecedent and consequent.

They assume that a conditional is considered true if the consequent is true in the context asserted by the antecedent and false if the consequent is false in such a context. This theory is consistent with a number of aspects of the everyday semantics of conditionals. For example, when untrained subjects negate a conditional, they usually negate the consequent and leave the antecedent unchanged (*if p then q* becomes *if p then not q*), presumably because the context (defined by the antecedent) is assumed to remain constant. This also explains why people often judge conditionals to be irrelevant in situations in which the antecedent is false (although logically, such conditionals are true) and why they only assert conditionals if there is some reason for relatinp the antecedent and consequent (e.g., people do not say, "If grass is green then the earth is round" even though it is logically true).

Of course, Johnson-Laird & Byrne go into much greater detail, describing the specific models they believe are built for each type of conditional, making explicit what they mean by all the terms in the summary above (e.g., "exhaustively represented"), and showing how this theory accounts for other aspects of the way people understand and use conditionals. Regardless of the details however, it is clear that this theory does not rely on validation. All three points in the summary above relate to how conditionals are interpreted during comprehension — they make no reference to the attempt to falsify conclusions by searching for alternative models.

I already mentioned that it is not surprising that in explaining the indeterminate interpretation of propositions, Johnson-Laird & Byrne made no reference to validation. Similarly, it is only natural that the mental model theory for the meaning of conditionals should be based on assumptions about comprehension rather than on assumptions about validation. But in the rest of Chapter 4, the authors go on to address how people *reason* with conditionals — the paradoxes of implication, the empirical results surrounding Wason's selection task, and the suppression of valid conditional inferences. If validation is central in deduction, then one would certainly expect it to play a role in explaining these kinds of results. As we will see, however, even these explanations do not depend on it.

The paradoxes of implication

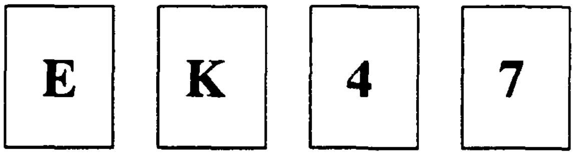
First, consider what Johnson-Laird & Byrne call the paradoxes of implication. As long as its antecedent is false, a conditional is guaranteed to be valid regardless of its consequent. For example, "if grass is red then Bush is president" is a logically valid proposition (since grass is not red). Similarly, as long as its consequent is true, a conditional with any antecedent whatsoever is guaranteed to be valid ("if I am president then grass is green"). But intuitively, neither proposition seems valid even though logically they are. Johnson-Laird & Byrne argue that the reason is that these propositions throw away semantic information. That is, there are more states of affairs in which the conditional is true than in which the negated antecedent (in the first case) or consequent (in the second) are true. Hence, "grass is not red" is more constraining than "if grass is red then Bush is president" and "grass is green" is more constraining than "if I am president then grass is green" and people would never use the conditional in such cases. But once again, notice that the search for alternative models does not play a role in accounting for these paradoxes.

Wason's selection task

Wason's selection task has attracted more attention than any other reasoning task (Wason, 1966). The task is shown in Figure 8-2. Subjects are shown four cards and are told that each card has a letter on one side and a number on the other. The upper (visible) faces of the cards contain the symbols E, K, 4, and 7. The subjects are then presented with a conditional proposition such as the following:

If a card has a vowel on one side,
then it has an even number on the other.

Their task is to indicate those cards (and only those cards) that *must* be turned over in order to decide whether the rule is true or false.



Which cards must be turned over to determine whether this rule is true:

If a card has a vowel on one side, then it has an even number on the other.

Figure 8-2: Wason's selection task.

One of the reasons that Wason's selection task has attracted so much attention is that it is so difficult. The correct behavior is to select the E and 7 cards and yet the vast majority of subjects either select only the E card, or the E and 4 cards. Performance can be facilitated by using certain types of realistic content. For example, if a conditional such as the following is used:

```
If someone is drinking beer then they must be over 18
```

and if one side of the cards specifies the age of a person in a bar while the other side specifies whether that person is drinking beer or a non-alcoholic drink, then subjects perform better.

Johnson-Laird & Byrne summarize the model theory of the selection task as follows (p. 79):

1. The subjects consider only those cards that are explicitly represented in their models of the rule.
2. They then select those cards for which the hidden value could have a bearing on the truth or falsity of the rule.

The authors go on to derive all the major regularities in the selection task from these simple assumptions.

Once again, though, note that neither assumption makes reference to the search for alternative models. So does the Johnson-Laird & Byrne theory constitute a verbal reasoning account of behavior on the selection task? Their explanations of the findings that I have previously discussed were all based solely on assumptions about comprehension and description, and could thus be clearly characterized as verbal reasoning. The same cannot be said for Wason's selection task. Neither of the assumptions above involves how subjects comprehend the problem statement or how they describe their mental model (though such assumptions also play a major role in their explanations), but rather what hypotheses (about the backs of cards) they will entertain and how they will test those hypotheses. As Evans (1982) put it, Wason's selection task "is not *simply* a deductive reasoning problem", but a "meta-inference task":

> Subjects are not simply required to draw or assess immediate inferences. Rather, they are invited to entertain alternative hypotheses with respect to the truth or falsity of a rule, and asked to test these hypotheses. (p. 157)

Johnson-Laird & Byrne (1991) make a similar point:

> The selection task calls for more than a deduction: subjects have to explore different possibilities, to deduce their consequences for the truth or falsity of a the rule, and on this basis to determine which cards to select. (p. 75)

It should not be surprising then that the selection task requires more than verbal reasoning. The reason is that the task requires knowledge or strategies that are not explicitly provided in the problem statement. As a result, the processes of comprehension and description, though important, are not sufficient. More generally, verbal reasoning will not generalize to any task that requires knowledge that is not available in the problem statement (specifically in a verbal format), and this provides an important characterization of the scope of verbal reasoning as a theory. Interestingly, while purely deductive tasks almost always satisfy this criterion, some non-deductive tasks do as well. To take a trivial example, consider a modified categorical syllogism in which the task is to generate a conclusion that is possible (rather than deductively valid) given the premises. Such a task does not require any knowledge outside of the problem statement and yet it is clearly not a deductive reasoning task. A theory based on verbal reasoning (e.g., a version of VR modified to reflect the different demands of the task) would apply quite naturally in such situations.

The suppression of valid deductions

Byrne (1986, 1989) has found that under certain circumstances, even the most natural inferences (e.g., modus ponens) can be suppressed. She used sets of premises such as the following:

```
If she meets her friends then she will go to a play.
If she has enough money then she will go to a play.
```

When subjects are presented with such premises and only one of the antecedents ("she meets her friends"), they tend *not* to make the valid modus ponens inference ("she will go to a play"). Johnson-Laird & Byrne argue that under these conditions, comprehension delivers one explicit model in which both antecedents and the consequent are true and one model with no explicit content. Since only one of the antecedents is asserted, the first model is rejected in favor of the second and no modus ponens inference is made. Like almost all of the previous explanations, this account makes no reference to validation, but is derived from assumptions about the models that comprehension delivers.

The spontaneous use of conditional descriptions

The final finding that Johnson-Laird & Byrne address in Chapter 4 involves how conditionals are used in descriptions. Byrne & Johnson-Laird (1990a, 1990b) presented subjects with factual sentences such as the following:

```
Laura has an essay to write.
The library stays open.
Laura studies late in the library.
```

When asked to combine them into a single sentence, subjects used conditionals on only

2% of trials — they tended to used connectives such as "and" or "when" instead. Johnson-Laird & Byrne assume that factual statements such as those above lead to a single explicit model in which all the assertions are true. According to the model theory, the representation of conditionals includes a model with no explicit content and so conditionals would not be used to summarize such statements. However, modal assertions such as "Laura *can* study late in the library" *are* predicted to lead to the creation of such implicit models and so conditionals should be used more frequently. Consistent with this prediction, conditionals appeared in 36% of such trials (compared with 2% on the other trials). In this case, the critical assumptions are about description — how will different types of models be described (specifically, when will conditionals be used in describing them). So this explanation is another example of a verbal reasoning account. Validation clearly does not play a role.

8.4. Reasoning about relations

Chapter 5 of *Deduction* begins with a review of three-term series problems (a.k.a. linear syllogisms) such as the following:

```
John is taller than Bill.
Mary is shorter than Bill.
Therefore, John is taller than Mary.
```

A major issue in the study of such tasks has been whether behavior can best be explained in terms of linguistic factors (Clark, 1969) or by assuming the construction of a spatial array (Huttenlocher, 1968). Johnson-Laird & Byrne consider Clark's linguistic account to be an example of a rule theory (that uses context-specific rules) while they consider Huttenlocher's imagery account to be a model theory. They argue that these tasks do not have enough structure to distinguish these different accounts and so they turn their attention to two-dimensional spatial deductions such as those shown in Figure 8-3. They show that for such tasks, theories based on rules vs. models make different predictions and that empirical results support model theory rather than rule theory.

For example, Hagert's rule theory (1983, 1984) predicts that problems I and II in the figure should be equally hard since the formal derivations of the conclusions contain the same number of steps. In contrast, model theory predicts that problem I should be easier because it is consistent with only one model of the premises:

```
C  B  A
D  E
```

while problem II is consistent with two:

```
C  A  B        A  C  B
D     E           D  E
```

Problem I	**Problem II**
A is on the right of B	B is on the right of A
C is on the left of B	C is on the left of B
D is in front of C	D is in front of C
E is in front of B	E is in front of B
What is the relation between D and E?	What is the relation between D and E?
Problem III	**Problem IV**
B is on the right of A	A is on the right of B
C is on the left of B	C is on the left of B
D is in front of C	D is in front of C
E is in front of A	E is in front of A
What is the relation between D and E?	What is the relation between D and E?

Figure 8-3: Two-dimensional spatial reasoning tasks.

While this prediction depends on the number of models, it nevertheless does not rely on validation. The reason is that the two models for problem II both support the same conclusion about D and E — that D is to the left of E. The second model does not falsify the initial conclusion, it confirms it. Clearly, then, this prediction is not based on searching for an alternative model that falsifies the putative conclusion.

There is one prediction in Chapter 5 that does make reference to validation, however. Problem III in Figure 8-3 is predicted to be harder than both problems I and II precisely because it requires validation to get it right while the other two problems do not. An initial model of problem III such as:

```
C    A    B
D    E
```

leads to the conclusion "D is left of E". But this conclusion can be falsified by considering the alternative model:

```
A    C    B
E    D
```

In keeping with this prediction, only 18% of subjects get tasks like problem III correct (compared with 61% and 50% for types I and II above). We seem finally to have an accurate prediction that critically depends on validation for its derivation. But notice that according to the model theory, the reason subjects get this task wrong is because they *fail* to successfully search for an alternative model. That is, according to the theory itself, subjects do *not* successfully apply the validation stage to these problems more than 80% of the time. And a substantial proportion of even these subjects may not be doing so. If they simply noticed the ambiguity of the premises during the initial encoding (as the authors assume they do for problem II above), then they might very well respond "no valid conclusion" without ever having constructed alternative models.

Finally, Johnson-Laird & Byrne predict that problem IV in Figure 8-3 should be relatively easy since it only supports one model:

```
C   B   A
D       E
```

This prediction is indeed borne out in the data (70% of subjects get this task correct). But again, the explanation makes no reference to validation.

To summarize our analysis of relational reasoning, only one of the model theory's explanations of empirical findings makes any reference to validation. And in explaining that finding, the model theory itself assumes that subjects rarely (less than 20% of the time) succeed in applying the strategy. Furthermore, even the behavior of those subjects who appear to validate can be explained more simply assuming they do not.

8.5. Categorical syllogisms

Having devoted the majority of this dissertation to showing that behavior on categorical syllogisms can best be explained in terms of verbal reasoning, I will not do so again here. Johnson-Laird & Byrne do, however, discuss two issues that VR has not addressed and so it is worth considering whether either of them provides strong evidence in favor of validation. The first involves subjects' memory of their responses while the second involves the effects of using "only" as the quantifier.

Byrne & Johnson-Laird (1989) had subjects solve a set of 16 syllogisms and subsequently asked them to choose the response they had given from a set of four alternatives. The alternatives never included "no valid conclusion" as a choice despite the fact that this was the correct answer on half of the problems and many subjects had, in fact, used it as their response. On 74% of the trials in which subjects had correctly responded "no valid conclusion", they chose the response that was consistent with considering only a single model. The authors suggest that this behavior can best be explained by assuming that the subjects initially considered that response, but rejected it when they constructed a falsifying model — they see it as evidence in favor of validation. But this behavior can be explained without the need for validation by assuming that the subjects are simply solving these problems again. Johnson-Laird and Byrne acknowledge as much:

> [The possibility] that they were reasoning from the premises once again ... cannot be eliminated. (p. 127)

The premise *Only the a's are b's* is logically equivalent to *All the b's are a's*, but Johnson-Laird & Byrne (1989) assume that *Only the a's are b's* leads to the explicit representation of negative information (specifically that anything that is *not* an *a* is also *not* a *b*) whereas *All the b's are a's* does not. They then argue that a model theory, augmented with this assumption can explain a variety of empirical results.

The first empirical finding they address is that problems using "only" are reliably harder than those using "all" (26% vs. 46% correct in Johnson-Laird & Byrne, 1989). They argue that the reason is "because the [initial] model for 'only' is more complex than the one for 'all'" (p. 129) — a verbal reasoning account based on the results of comprehension rather than on validation. Second, problems that did not require validation were much easier (55% correct) than those that did (and had a valid conclusion) (15% correct). As in the case of relational reasoning, the model theory itself assumes that subjects successfully validate their conclusions less than 20% of the time. Also, since valid conclusions are guaranteed to be true in all models of the premises, they are guaranteed to be true in any initial model the subjects may construct. So the subjects that did get these problems correct may just have fortuitously generated the valid conclusion from the initial model without ever having constructed alternative models. In fact, this response could be more than fortuitous. If the subject is able to incorporate the demands of logical necessity into the encoding and reencoding processes (so that they encode all necessary knowledge but nothing unwarranted), then conclusions based on the initial model could be guaranteed to be valid. Third, when both premises contained the quantifier "only", just 16% of conclusions used it — a result which runs counter to the standard atmosphere effect in syllogistic reasoning. Johnson-Laird & Byrne suggest that since these models are also consistent with "all" conclusions, subjects prefer using that quantifier since it does not require processing negative information. Of course, these are assumptions about comprehension and description rather than validation and so this too is a verbal reasoning explanation. Finally, when Johnson-Laird & Byrne presented subjects with premises such as:

```
All authors are bankers.
Mark is an author.
```

and:

```
Only bankers are authors.
Mark is not a banker.
```

they found an interaction between the quantifier ("all" or "only") and the polarity or quality of the second premise (whether or not it was negated). Specifically, when "all" was used, subjects performed significantly worse if the second premise was negated than if it was not (73% vs. 96% correct). But if "only" was used, no such difference was found (86% vs. 90% correct). Johnson-Laird & Byrne argue that the reason is that "only", but not "all", leads to the explicit representation of negative information (about non-bankers). Consequently, inferring that Mark is not an author (in the second example above) should be just as easy as inferring that an author is a banker — both are supported in the initial model. In contrast, the initial model for "all" only supports the inference from an affirmative premise and so making inferences from negated premises should be harder. Again though, the critical assumptions are about the initial model delivered by comprehension, not about validation.

8.6. Reasoning with multiple quantifiers

In Chapter 7 of *Deduction*, Johnson-Laird & Byrne turn to multiply-quantified deductions such as those shown in Figure 8-4. They present results from three experiments (from Johnson-Laird, Byrne, & Tabossi, 1989) that were designed to distinguish mental model theory from theories based on formal rules. Their data run counter to rule theories but are consistent with model theory.

A one-model problem

None of the painters is in the same place as any of the musicians.
All of the musicians are in the same place as all of the authors.
Therefore, none of the painters is in the same place as any of the authors.

A multiple-model problem

None of the painters is in the same place as any of the musicians.
All of the musicians are in the same place as some of the authors.
Therefore, none of the painters is in the same place as some of the authors.

Figure 8-4: Multiply-quantified deductive reasoning tasks.

The main result in all three experiments is the same: problems that are consistent with multiple models are harder than those that are consistent with only one. The authors make explicit reference to validation in explaining this result. To get the multiple model problems correct, subjects must consider alternative models that falsify their initial conclusions. Since there are no such alternative models to consider for one-model problems, these tasks place less of a load on working memory and are consequently easier.

Can these results be explained in terms of verbal reasoning without making reference to validation? Consideration of a number of points illustrates that they can. First of all, as in previous cases, very few subjects correctly solved the problems that are assumed to require validation. On valid multiple-model problems in the three experiments (those that require validation), only 13%, 16%, and 23% of responses were correct. So once again, according to the model theory itself, subjects do not successfully validate their conclusions very often. Subjects performed fairly well on *invalid* multiple-model problems (those with no valid conclusion), correctly solving 50%, 40%, and 23% of these tasks in the three experiments. But this result cannot be attributed to the use of validation since that would not predict the large difference in performance between valid and invalid multiple-model problems (if subjects can successfully validate on the invalid problems, they should also be able to validate on the valid problems).

Verbal reasoning provides a natural account of the difference between one-model and multiple-model problems. The reason one-model problems are easy is because any initial model of the premises will only support valid conclusions (otherwise they could be falsified by an alternative model and it would not be a one-model problem). In contrast, the initial models for multiple-model problems can support invalid conclusions (that is why they can be falsified by alternative models) and so they should be harder. Furthermore, even the rare cases that seem to suggest validation (correct responses to valid multiple-model problems) can be explained more simply without it. Rather than constructing a sequence of alternative models and finding a conclusion that is true in all of them, subjects may just be fortuitously generating the valid conclusion from the initial model. After all, since the conclusion is valid, it is guaranteed to be supported in any model of the premises, including the one that is constructed initially. And if the initial encoding and reencoding of the problem is adapted to the needs of logical necessity (incorporating all valid information but nothing unwarranted), then such conclusions need not be lucky — the subject could get the answer right for the right reasons. In any case, one need not assume validation to account for this behavior. Similarly, correct NVC responses to invalid syllogisms do not necessarily imply the search for alternative models — subjects may either be unable to construct a model that supports any legal conclusion or they may notice some ambiguity during encoding and decide to respond NVC rather than risk a conclusion that they realize may not be valid.

More generally, the above argument shows that for *any* task, a difference in difficulty that is predicted based on the use of validation (one vs. many models) can be explained more simply by a verbal reasoning account. Subjects find multiple-model problems more difficult, not because they search through a sequence a alternative models, but because the initial model they construct supports invalid conclusions. By definition, the initial model for one-model problems cannot support invalid conclusions and so these problems are easier.

8.7. Meta-deduction

The final reasoning domain that Johnson-Laird & Byrne address is meta-deduction. They distinguish between two types of meta-deduction: *meta-logical* reasoning and *meta-cognitive* reasoning and I will consider each in turn.

Meta-logical reasoning problems make explicit reference to truth and falsity. For example, consider the following "knight-and-knave" puzzle:

> Knights always tell the truth while knaves always lie. Lancelot says, "I am a knave and so is Gawain". Gawain says, "Lancelot is a knave". What are Lancelot and Gawain?

The correct answer is that Lancelot is a knave (he is lying when he says Gawain is a knave and hence the conjunction is also a lie) while Gawain is a knight.

Johnson-Laird & Byrne propose that subjects employ certain meta-logical strategies in addition to their standard processes for constructing a model, generating a conclusion from a model, and searching for alternative models. They summarize these strategies as follows:

1. Simple chain: assume that the assertor in the first premise tells the truth, and follow up the consequences, but abandon the procedure if it becomes necessary to follow up disjunctive consequences. Assume that the assertor in the first premise is lying and do likewise.

2. Circular: if a premise is circular, follow up the immediate consequences of assuming that it is true, and then follow up the immediate consequences of assuming that it is false.

3. Hypothesize-and-match: if the assumption that the first assertor A is telling the truth leads to a contradiction, then attempt to match -A with the content of other assertions, and so on.

4. Same-assertion-and-match: if two assertions make the same claim, and a third assertor, C, assigns the two assertors to different types, or *vice versa*, then attempt to match -C with the content of other assertions, and so on.

Based on these strategies and the basic model theory, the authors go on to derive three predictions about meta-logical reasoning. For our purposes, the specific details of the theory (e.g., exactly how the above strategies work) are not important. What is important is the role that validation plays in the explanations. And as we will see, none of these three predictions depend on validation.

The first prediction is that problems that can solved using one of the four strategies above will be easier than problems that cannot. Consistent with this prediction, problems for which one of the above strategies was sufficient were correctly solved 28% of the time, compared with only 14% correct responses on other problems. Clearly though, this prediction is based on the specific meta-logical strategies above, none of which make any reference to validation. Johnson-Laird & Byrne state the second prediction as follows: "the difficulty of a problem will depend on the number of clauses that it is necessary to use in order to solve the problem" (p. 160). They assume that additional clauses put extra strain on working memory and make the problem more difficult. They go on to spell out what they mean by using a clause (basically making one inferential step), but regardless of the details, it is clear that this prediction does not depend on searching for alternative models. Indeed, they acknowledge that it really does not depend on the basic model theory at all: "the prediction is almost independent of the processing theory that we have proposed, and is likely to be made by any sensible analysis of meta-logical problems" (p. 160). The third prediction is that the hypothesis that an assertion is true should be easier to process than the hypothesis that an assertion is false. The reason is that negation causes problems. Again, it is clear that this prediction does not depend on validation at all.

Johnson-Laird & Byrne go on to address a second kind of meta-deduction that they refer to as meta-cognitive reasoning. These tasks involve deducing what someone else could have deduced. For example, consider the following problem:

> Three wise men who were perfect logicians were arrested by the Emperor on suspicion of subversion. He put them to the following test. The three men were lined up in a queue facing in the same direction, and a hat was placed on the head of each of them. The men could not see their own hats, but the man at the back of the queue (A) could see the two hats in front of him, the man in the middle (B) could see the one hat in front of him, and the man at the front (C) could see no hat. The Emperor said: "If one of you can tell me the colour of your own hat, I will set all three of you free. There are three white hats and two black ones from which your hats have been drawn. I will now ask each of you if he can tell me the colour of his hat. You may answer only 'yes', 'no', or 'I don't know'". A who could see the two hats in front of him said, "I don't know". B heard A's answer and said, "I don't know". C heard the two previous answers. What was C's answer? (Johnson-Laird & Byrne, 1991, p. 162)

It turns out that C can deduce the color of his hat (it must be white) based on the answers of A and B (who are perfect logicians). A could not have seen two black hats since otherwise he could have deduced that his own hat was white. Consequently, one (or both) of B and C are wearing a white hat. So if C's hat were black, then B could have deduced that his own hat was white. Since he did not, C knows that his hat must be white.

The only empirical result about tasks like this that Johnson-Laird & Byrne address is the fact that they are difficult. They propose that there are two major sources of difficulty: working memory load and the lack of an appropriate strategy. The task above requires constructing models of C's models which are in turn models of A's and B's models of the situation. Keeping track of all this information undoubtedly strains the limits of working memory. And the appropriate strategy for solving the problem is far from obvious at first glance. For our purposes, however, the question is whether this explanation of the difficulty of these tasks depends on validation. And since neither the amount of information that must be represented nor the lack of an appropriate strategy involve the search for alternative models, the answer is clearly no.

This fact does not imply that a verbal reasoning account will be able to account for these phenomena however. As I emphasized in the discussion of Wason's selection task, verbal reasoning will not generalize to tasks that require knowledge that is not available in the problem statement in a verbal format. And the knowledge incorporated in the strategies proposed by Johnson-Laird & Byrne is obviously not available in the problem statement. As a result, it is doubtful that the processes of comprehension and description would be sufficient to account for behavior on this task.

8.8. Conclusion

I have now reviewed all six of the deductive reasoning domains addressed by Johnson-Laird & Byrne: propositional reasoning, conditional reasoning, relational reasoning, syllogistic reasoning, reasoning with multiple quantifiers, and meta-deduction. In some cases (propositional reasoning and most of conditional reasoning), I showed that the model theory's explanations really only depended on comprehension and description and could thus be reinterpreted as verbal reasoning. In other cases (syllogistic reasoning and reasoning with multiple quantifiers), I demonstrated that the model theory itself assumed validation was rare and showed that alternative accounts based on verbal reasoning alone could account for the behavior more simply. In short, these analyses demonstrated that verbal reasoning can account for behavior across a wide variety of reasoning tasks — it has extensive breadth.

But these analyses also showed that verbal reasoning cannot account for behavior that depends on knowledge that is not provided verbally in the problem statement. Johnson-Laird & Byrne's explanations of behavior on Wason's selection task and in meta-deduction depended on assumptions beyond both model theory *and* verbal reasoning (specifically, about the types of strategies that subjects employ), probably because both tasks involve meta-inference. Obviously, one would not expect linguistic processes alone to be sufficient to account for such behavior. But on all the standard deductive reasoning tasks (that do not involve meta-inference), verbal reasoning provides accurate accounts of human behavior.

Chapter 9
Conclusion

I began this thesis with a very general question: what is the nature of the cognitive processes that people apply when they reason? This question is probably too vague to answer precisely (e.g., what counts as reasoning?), so I focused on human behavior on certain well-defined reasoning tasks. The first task I examined was the categorical syllogism. I described VR, a computational model of syllogistic reasoning, and showed that it provides the most detailed and accurate account of behavior on this task to date. It simulates all the standard phenomena that have been discovered, it makes a number of empirically accurate novel predictions, and perhaps most importantly, it models the behavior of individual subjects with remarkable accuracy.

Previous theories of syllogistic reasoning have shared the same basic structure at the most general level — encode the problem statement into some internal representation, apply certain general-purpose reasoning processes on that representation (e.g., the search for alternative models, the application of formal or content-specific rules of inference), and then decode the result. But VR presents a very different view. The central processes in VR's behavior are not sophisticated general-purpose reasoning skills, but just processes for encoding and reencoding the verbal problem statement.

VR's structure led to the general idea of verbal reasoning — the deployment of linguistic processes according to and in order to satisfy the demands of a reasoning task. VR's empirical success demonstrated that verbal reasoning provided an accurate characterization of the behavior of untrained subjects on categorical syllogisms. A natural question was whether verbal reasoning would generalize to any other tasks and that is what I investigated next.

I attacked this question in a rather unorthodox way. Rather than developing verbal reasoning theories for behavior on other tasks (the number of which would be severely limited by time constraints), I exploited the extensive effort that has already gone into demonstrating the generality of an alternative account — mental model theory. This theory assumes the people reason by encoding the problem statement into a mental model of the situation being described (comprehension), generating conclusions based on that model (description), and searching for alternative models that falsify their putative conclusions (validation). Over the last ten years, Johnson-Laird and his colleagues have showed that this theory can account for human behavior on all the standard tasks used in studying deductive reasoning. I was able to show that in most cases, the validation process did not provide any explanatory power in their accounts. Even in the cases in which it did, I was able to provide alternative accounts that did not depend on validation. Since without validation, mental model theory reduces to the purely linguistic processes of comprehension and description, these analyses showed that verbal reasoning could account for behavior on all the standard deductive reasoning tasks.

This analysis also provided insight into the scope of verbal reasoning. The mental model accounts of two of the tasks critically depended on assumptions about the meta-logical strategies that subjects employed. In analyzing these tasks it became clear that verbal reasoning could not extend to such behavior because it depended on knowledge that was not provided verbally in the problem statement — an important characterization of the limits of verbal reasoning as a theory.

References

Anderson, J. R. (1990). *The Adaptive Character of Thought.* Hillsdale: Lawrence Erlbaum Associates.

Begg, I. and Denny, P. (1969). Empirical reconciliation of atmosphere and conversion interpretations of syllogistic reasoning errors. *Journal of Experimental Psychology, 81*, 351-354.

Begg, I. and Harris, G. (1982). On the interpretation of syllogisms. *Journal of Verbal Learning and Verbal Behavior, 21*(5), 595-620.

Braine, M.D.S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review, 85*(1), 1-21.

Byrne, R. M. J. (1986). The contextual nature of conditional reasoning. PhD thesis, University of Dublin.

Byrne, R. M. J. (1989). Suppressing valid inferences with conditionals. *Cognition, 31*, 61-83.

Byrne, R. M. J. & Johnson-Laird, P. N. (1989). Re-constructing inferences. Unpublished manuscript, University of Wales College of Cardiff.

Byrne, R. M. J. & Johnson-Laird, P. N. (1990). The use of propositional connectives. Manuscript submitted for publication, University of Wales College of Cardiff.

Byrne, R. M. J. & Johnson-Laird, P. N. (1990). Models and deduction. In K. Gilhooly, M. T. G. Keane, R. Logie, & G. Erdos (Eds.), *Lines of thought: Reflections on the psychology of thinking (Vol. 1).* Wiley.

Ceraso, J. and Provitera, A. (1971). Sources of error in syllogistic reasoning. *Cognitive Psychology, 2*, 400-410.

Chapman, L. J. and Chapman, J. P. (1959). Atmosphere effect re-examined. *Journal of Experimental Psychology, 58*(3), 220-226.

Cheng, P. W. & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17*(4), 391-416.

Clark, H. H. (1969). Linguistic processes in deductive reasoning. *Psychological Review, 76*(4), 387-404.

Dickstein, L. S. (1975). Effects of instruction and premise order on errors in syllogistic reasoning. *Journal of Experimental Psychology: Human Learning and Memory, 104*, 376-384.

Dickstein, L.S. (1978). The effect of figure on syllogistic reasoning. *Memory and Cognition, 6*(1), 76-83.

Erickson, J.R. (1974). A set analysis theory of behavior in formal syllogistic reasoning tasks. In Solso, R. (Eds.), *Theories in cognitive psychology: the Loyola symposium.* Lawrence Erlbaum Associates.

Evans, J. St. B. T. (1982). *The psychology of deductive reasoning.* London: Routledge & Kegan Paul.

Evans, J. St. B. T. & Newstead, S. E. (1980). A study of disjunctive reasoning. *Psychological Research, 41*, 373-388.

Evans, J. St. B.T., Barston, J. L., and Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition, 11*, 295-306.

Fisher, D.L. (1981). A three-factor model of syllogistic reasoning: the study of isolable stages. *Memory and Cognition, 9*(5), 496-514.

Gonzalez-Marques, J. (1985). La influencia de materiales no emocionales en la solucion de silogismos categoricos. (The effects of nonemotional stimulus materials on the solution of categorical syllogisms). *Informes-de-Psicologia, 4*(3), 183-198.

Guyote, M.J. and Sternberg, R.J. (1981). A transitive-chain theory of syllogistic reasoning. *Cognitive Psychology, 13*(4), 461-525.

Hagert, G.. (1983). Report of the Uppsala Programming methodology and artificial intelligence laboratory.

Hagert, G. (1984). Modeling mental models: Experiments in cognitive modeling of spatial reasoning. In T. O'Shea (Eds.), *Advances in artificial intelligence.* Amsterdam: North Holland.

Henle, M. (1962). On the relation between logic and thinking. *Psychological Review, 69,* 366-378.

Huttenlocher, J. (1968). Constructing spatial images: A strategy in reasoning. *Psychological Review, 75*(6), 550-560.

Inder, R. (1986). Modeling syllogistic reasoning using simple mental models. In Cohn, A. G. and Thomas, J. R. (Eds.), *Artificial Intelligence and Its Applications.* New York, New York: Wiley.

Inder, R. (1987). The Computer Simulation of Syllogism Solving using Restricted Mental Models. Unpublished doctoral dissertation.

Janis, I. L. and Frick, F. (1943). The relationship between attitudes towards conclusions and errors in judging logical validity of syllogisms. *Journal of Experimental Psychology, 33,* 73-77.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Harvard University Press.

Johnson-Laird, P. N. and Byrne, R.M.J. (1991). *Deduction.* Lawrence Erlbaum Associates.

Johnson-Laird, P.N. and Bara, B.G. (1984). Syllogistic Inference. *Cognition, 16,* 1-61.

Johnson-Laird, P.N. and Byrne, R. M. J. (1989). *Only* reasoning. *Journal of Memory and Language, 28,* 313-330.

Johnson-Laird, P.N. and Steedman, M. (1978). The psychology of syllogisms. *Cognitive Psychology, 10,* 64-99.

Johnson-Laird, P.N., Byrne, R. M. J., & Tabossi, P. (1989). Reasoning by model: The case of multiple quantification. *Psychological Review, 96,* 658-673.

Lee, J. R. (1983). Johnson-Laird's mental models: two problems. Discussion paper, School of Epistemics, University of Edinburgh.

Lehman, J. Fain, Newell, A., Polk, T., and Lewis, R. L. (in press). The Role of Language in Cognition: A Computational Inquiry. In Harman, G. (Eds.), *Conceptions of the Human Mind.* Lawrence Erlbaum Associates, Inc.

Manktelow, K. I. (1980). The role of content in reasoning. Unpublished PhD thesis, Plymouth Polytechnic.

Morgan, J. I. B. and Morton, J. T. (1944). The distortions of syllogistic reasoning produced by personal connections. *Journal of Social Psychology, 20,* 39-59.

Newell, A. (1990). *Unified Theories of Cognition.* Cambridge, Massachusetts: Harvard University Press.

Oakhill, J. V. and Johnson-Laird, P. N. (1985). The effects of belief on the spontaneous production of syllogistic conclusions. *Quarterly Journal of Experimental Psychology, 37A,* 553-569.

Oakhill, J. V., Johnson-Laird, P. N., and Garnham, A. (1989). Believability and syllogistic reasoning. *Cognition, 31,* 117-140.

Polk, T.A. & Newell, A. (1992). A verbal reasoning theory for categorical syllogisms. Unpublished manuscript, Carnegie Mellon University.

Polk, T. A. and Newell, A. (1988). Modeling human syllogistic reasoning in Soar. *Proceedings of the Annual Conference of the Cognitive Science Society, 10th,* 181-187.

Polk, T.A., Newell, A., and VanLehn, K. (1992). Analysis of symbolic parameter models (ASPM): a new model-fitting technique for the cognitive sciences. Forthcoming.

Polk, T.A., Newell, A., and Lewis, R.L. (1989). Toward a unified theory of immediate reasoning in Soar. *Proceedings of the Annual Conference of the Cognitive Science Society, 11th*, 506-513.

Revlis, R. (1975a). Syllogistic reasoning: logical decisions from a complex data base. In Falmagne, R.J. (Eds.), *Reasoning: Representation and Process*. Lawrence Erlbaum Associates.

Revlis, R. (1975b). Two models of syllogistic reasoning: Feature selection and conv^rsion. *Journal of Verbal Learning and Verbal Behavior, 14*, 180-195.

Revlis, R., Ammerman, K., Petersen, K. and Leirer, V. (1978). Category Relations and Syllogistic Reasoning. *Journal of Educational Psychology, 70*(4), 613-625.

Rips, L.J. (1983). Cognitive processes in propositional reasoning. *Psychological Review, 90*(1), 38-71.

Sells, S. B. (1936). The atmosphere effect: an experimental study of reasoning. *A:chives of Psychology,* Vol. *200.*

Siegler, R. S. (1987). The perils of averaging data over strategies: an example from children's addition. *Journal of Experimental Psychology: General, 116*(3), 250-264.

Wason, P. C. (1966). Reasoning. In Foss, B. M. (Eds.), *New horizons in psychology*. Penguin.

Wilkins, M. C. (1928). The effect of changed material on the ability to do formal syllogistic reasoning. *Archives of Psychology,* Vol. *102.*

Woodworth, R. S. and Sells, S. B. (1935). An atmosphere effect in formal syllogistic reasoning. *Journal of Experimental Psychology, 18*, 451-460.